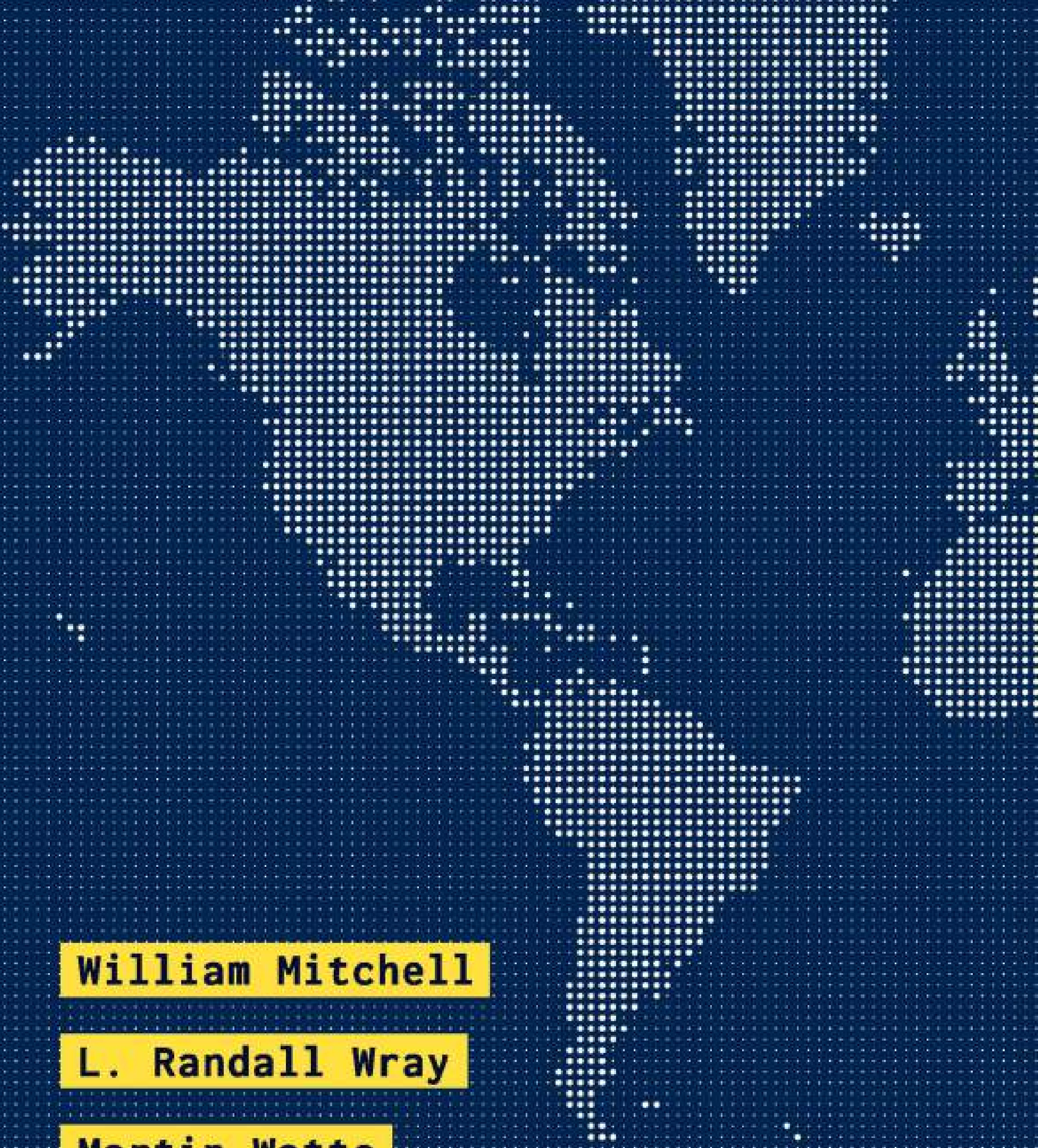


William Mitchell

L. Randall Wray

Martin Watts

MACROECONOMICS

Finally, a macroeconomics text that takes a modern-monetary approach to the macroeconomy. Finally, an alternative to the mainstream vision of the macro economy and what it says and, more importantly, fails to say, about the real world. The presentation of Keynes closely approximates his vision, a much-needed change from texts that present a "bastardized" view of Keynes' theory. The book is highly readable, the assumptions underlying the different models are presented clearly, the logic easy to follow, and the historical context serves to add relevance.

—**John P. Watkins,** *Westminster College, Utah, USA*

A complete, up-to-date and excellent account of macroeconomic theory. An excellent critique of the mainstream economic paradigm, it calls for more realistic approaches to human behaviour, and for heterodox ideas in economics. It presents macroeconomic methodology in the light of the 2007-08 crisis, and outlines how economics should proceed in the post-crash world. One of the most interesting and remarkable macroeconomics textbooks of the past few years.

—**Omar Feraboli,** *University of Dundee, UK*

Finally, a macro textbook that rejects neoclassical microfoundations as a basis for understanding how capitalism works. The authors replace maximizing individuals with social classes where capitalist firms exercising differential power largely determine economic outcomes.

—**Robert Chernomas,** *University of Manitoba, Canada*

The most progressive macroeconomics textbook on the market. Organized around a balance sheet view, the authors carefully examine the most important issues of our time: Why government as the sovereign issuer of currency is not financially constrained, how banks create deposits by making loans and how the central bank influences the economy by shifting interest rates. The authors are outstanding scholars of macroeconomics, and this textbook should enlighten the next generation of students. I recommend it wholeheartedly.

—**Dirk Ehnts,** *Europa-Universität Flensburg, Germany*

A uniquely well-structured, clear discussion of macroeconomics for the 21st century, contrasting a useful and realistic analysis of a modern monetary production economy with the flawed, unrealistic and outdated narrowly neoclassical approach which has failed so comprehensively in recent years. Quite simply the best economics textbook I have ever read, and one which should become a standard text in many universities and colleges in the years to come.

—**Steven Hail,** *University of Adelaide, Australia*

Supremely accessible...Unlike most other textbooks, it explains schools of thought, relates them to the history of economic ideas and provides historical and institutional detail. Students are invited to reflect on why economists disagree, what macroeconomics means, and the role of language and framing. One can only wish that all textbooks nurtured sensitivity to these topics in future economists. All in all, an exceptionally clear text for students new to macroeconomics and a good grounding in the current policy debates shown in the final part of the book.. It marries pluralism of ideas with a clear exposition of Modern Monetary Theory, all with institutional realism and analytical rigor - a major feat. Highly recommended!

—**Dirk J. Bezemer,** *University of Groningen, Netherlands*

Even some mainstream economists now acknowledge that the macroeconomics of the past thirty years represents a big step back. But they do not recognize that this is because it lacks a pluralist perspective, and in the meantime they have ignored the revolutionary potential of heterodox economics for reconstructing macroeconomics. This new book points out the right direction for the coming macroeconomics revolution.

—**Jia Genliang,** *Renmin University of China (Beijing), China*

This is a long awaited first-year university textbook where heterodox and mainstream approaches run side by side within a critical framework which avoids the pedagogical pitfall of forcing students to learn economics through mainstream lenses. Students and experienced economists alike can learn from this wonderful book.

—**Alberto Paloni**, *University of Glasgow, UK*

Macroeconomics provides an exceptionally comprehensive and balanced guide to macroeconomics from both orthodox and heterodox views (post-Keynesian, institutionalists, and Marxian). It covers important but rarely discussed topics at the introductory level such as measurement of inequality, endogenous money and Modern Money Theory (MMT), quantitative easing, negative interest rates policies, theory of effective demand, full employment policy, economic instability, and environmental sustainability. This is a truly important and innovative introduction to macroeconomics, not only for beginning students, but also even for professional economists who would like an introduction to some of these important subjects that orthodox economics neglect. Kudos to the authors!

—**Y.K. Kim**, *University of Massachusetts Boston, USA*

MACROECONOMICS

WILLIAM MITCHELL

L. RANDALL WRAY

MARTIN WATTS

macmillan international HIGHER EDUCATION

RED GLOBE PRESS

First published 2019 by
RED GLOBE PRESS

Red Globe Press in the UK is an imprint of Springer Nature Limited, registered in England, company number 785998, of 4 Crinan Street, London, N1 9XW.

Red Globe Press® is a registered trademark in the United States, the United Kingdom, Europe and other countries.

ISBN 978–1–137–61066–9 paperback

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources. Logging, pulping and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

A catalogue record for this book is available from the British Library.

A catalog record for this book is available from the Library of Congress.

CONTENTS

PART C: NATIONAL INCOME, OUTPUT AND EMPLOYMENT DETERMINATION

LIST OF FIGURES

LIST OF TABLES

LIST OF BOXES

ABOUT THE AUTHORS

William Mitchell

William Mitchell is Professor of Economics at the University of Newcastle, NSW, Australia, Director of the Centre of Full Employment and Equity (CofFEE) and a Docent Professor of Global Political Economy at the University of Helsinki, Finland.

William is the author of various books, the most recent being *Eurozone Dystopia* (Elgar, 2015) and *Reclaiming the State* (Pluto, 2017). He has published widely in refereed academic journals and books in the areas of macroeconomics, labour market studies, econometric modelling, regional economics and economic development. He has received regular competitive research grant support and has extensive experience as a consultant to the Australian Government, trade unions and community organisations, and several international organisations (including the European Commission; the International Labour Organisation and the Asian Development Bank). William maintains a high commitment to community activities, and regularly provides media commentary on economics. He is one of the founders of Modern Monetary Theory (MMT).

L. Randall Wray

L. Randall Wray is a Professor of Economics at Bard College and Senior Scholar at the Levy Economics Institute of Bard College, NY. He is a past-president of the Association for Institutionalist Thought (AFIT) and has served on the board of directors of the Association for Evolutionary Economics (AFEE). A student of Hyman P. Minsky, Randall has focused on monetary theory and policy, macroeconomics, financial instability, and employment policy. He is one of the founders of MMT.

Randall is the co-editor (with Jan Kregel) of the *Journal of Post Keynesian Economics*. His books include *Modern Money Theory: A Primer on Macroeconomics for Sovereign Monetary Systems* (Palgrave, 2012; second revised edition, 2015); *Understanding Modern Money: The Key to Full Employment and Price Stability* (Elgar, 1998) and *Money and Credit in Capitalist Economies: The Endogenous Money Approach* (Elgar, 1990). He edited *Theories of Money and Banking* (two volumes, Elgar, 2012) and *Credit and State Theories of Money: The Contributions of A. Mitchell Innes* (Elgar, 2004). He co-authored (with Eric Tymoigne) *The Rise and Fall of Money Manager Capitalism; Minsky's Half-Century: From World War Two to the Great Recession* (Routledge, 2013).

Randall's newest book is *Why Minsky Matters* (Princeton University Press, 2016), and he is currently editing an unfinished book manuscript by the late Hyman Minsky.

Martin Watts

Martin Watts is Emeritus Professor of Economics at the University of Newcastle, NSW, Australia, and Research Associate of the Centre of Full Employment and Equity (CofFEE). He was previously employed at Monash University (1975–90).

Martin has published widely in refereed academic journals and written many book chapters in the area of macroeconomics, the conceptualisation and measurement of segregation and spatial modelling. He has regularly received competitive research grant funding and has extensive teaching experience, primarily in labour economics and macroeconomics.

ABOUT THE BOOK

Macroeconomics has eight parts. In Part A *Introduction and Measurement*, we introduce students to the subject matter of macroeconomics, and how it differs from microeconomics (Chapter 1). We note that it is a highly contested discipline, and that macroeconomic reasoning can be blighted by the fallacy of composition. The importance of developing skills of critical thinking is emphasised (Chapter 2). In Chapter 3, we place capitalism in context by a brief overview of economic history in which its rise to prominence is explored. Every discipline has its own language in the form of concepts and theories that provide the basis for understanding, and not merely describing, the relevant phenomena. To this end, we develop some initial conceptual understanding of national accounts, the labour market and sectoral balances (Chapters 4–6). Concepts and theories can also be depicted and understood through the development of formal mathematical models. Some introductory mathematical material is provided in Chapter 7. Students need to recognise the importance of framing and language in learning macroeconomics (Chapter 8).

In Part B *Currency and Banking*, we explain why a fiat currency is valued and is acceptable in domestic transactions. The distinction between fixed and floating exchange rate regimes and their significance for the conduct of macroeconomic policy is explained. Students are provided with an understanding of how IOUs are created and extinguished (Chapter 9). The focus is money and banking in Chapter 10. The definitions of the money supply and financial assets are outlined. The important distinction between the MMT and orthodox representations of the process of credit creation by banks is highlighted. Also, students are introduced to simplified balance sheets, which provide important insights as to the operation of the financial system.

In Part C *National Income, Output and Employment Determination*, a number of models are outlined, beginning with the Classical system which still influences macroeconomic theory and policy today (Chapter 11). This is followed by Keynes' rebuttal of Classical theory due to major flaws in its analysis of both interest rate and employment determination (Chapter 12) and his demonstration that employment and output depend on expected effective demand (Chapter 13). The macroeconomic demand for labour is argued to be a derived demand and it is shown that macroeconomic equilibrium can be characterised by unemployment (Chapter 14). Part C concludes with the presentation of the real expenditure model (Chapter 15) and a detailed analysis of mark-up pricing theory which provides a rationale for firms acting as quantity adjusters in the short run in the real expenditure model.

In Part D *Unemployment and Inflation: Theory and Policy*, we first define inflation and go on to argue that it emanates from a conflict over the distribution of income. We highlight the deficiencies of the Quantity Theory of Money (Chapter 17). In Chapter 18 the early Phillips Curve debate is outlined, and this is followed by a critical analysis of the expectations augmented Phillips Curve which continues to have a profound influence on the conduct of macroeconomic policy in developed economies more than 40 years later. Students are also exposed to recent advances in the Phillips curve literature which include hysteresis and hence the importance of the duration of unemployment, and also the role of underemployment. Most policymakers continue to utilise a buffer stock of unemployment to counter inflationary pressure. Chapter 19 explores the merits of a Job Guarantee which is based on an employment buffer stock and is designed to achieve both full employment and price stability in concert with other macroeconomic policies.

In Part E *Economic Policy in an Open Economy*, we start with an introductory outline of the roles of the treasury and central bank. Typically, liquidity management by the central bank must accompany the operation of fiscal policy. We argue that a necessary condition for the capacity to conduct independent monetary policy is currency sovereignty. Also, the design of the taxation system should be motivated by equity and behavioural objectives and not revenue raising. In Chapter 21, the competing views about the conduct of fiscal policy associated with the deficit hawks, doves and owls are outlined. The crowding out arguments are presented and rejected. In addition, the alleged link between stimulatory fiscal policy and hyperinflation is assessed. We develop the concepts of fiscal space and fiscal sustainability in Chapter 22. A floating exchange rate is shown to maximise fiscal space, and we demonstrate that a government which operates with a sovereign currency can never face a crisis associated with public debt sustainability. Chapter 23 is devoted to an analysis of the operation of monetary policy by the central bank and its impact on the macroeconomy. The final chapter in Part E outlines the components of the balance of payments and their interrelationship. We make the distinction between nominal and real exchange rates. The real expenditure model (introduced in Chapter 15) is extended in Chapter 24 to incorporate foreign trade and the impact of changes in the exchange rate.

Part F explores *Economic Instability*, with the focus in Chapter 25 being the role of investment due to the operation of both the multiplier and the accelerator. The insights gained from the modelling of investment are examined in the context of policies to stabilise the economy. In Chapter 26, Marxist and other heterodox theories of crisis which are financial in their origins are presented, with a particular focus on Minsky's financial instability hypothesis, which has the counter-intuitive conclusion that remedial policies enacted by treasury and the central bank can further weaken the financial system, and make it more prone to crisis.

Part G is entitled *History of Macroeconomic Thought* and starts with an overview of the history of economic thought, commencing with Smith and then covering the neoclassicists of the late 19th century. A chronology is presented which shows when particular economic principles were developed and how they have been incorporated (or rejected) by later schools of thought. Chapter 28 is devoted to a thorough examination of a very influential body of theoretical work: IS-LM analysis. Policy options are analysed within the IS-LM framework, and also the macroeconomics consequences of wage/price flexibility. Finally, the limitations of the IS-LM framework are documented. In Chapter 29 modern schools of economic thought are outlined and assessed. In particular, New Classical, Real Business Cycle and New Keynesian perspectives are developed, along with modern heterodox approaches.

The final three chapters comprise Part H *Contemporary Policy Debates*. In Chapter 30, we outline the new monetary consensus in macroeconomics. This body of economic thought can be viewed as the post-Global Financial Crisis consolidation of the modern orthodox schools of thought. In the following chapter, five contemporary policy debates are analysed through an MMT lens. These are (i) Ageing, Social Security, and the Intergenerational Debate; (ii) the Twin Deficits Hypothesis; (iii) Balance of Payments Constraints and Currency Crises; (iv) Fixed versus Flexible Exchange Rates: Optimal Currency Areas, the Bancor, or Floating Rates; and (v) Environmental Sustainability and Economic Growth. In Chapter 32, the key deficiencies of mainstream macroeconomics are identified; shortcomings that meant that its supporters did not foresee the Global Financial Crisis (GFC). By contrast, we outline the reasons why MMT advocates both anticipated the GFC and recognised the design faults of the European Monetary Union. In the final chapter, the key building blocks of a robust macroeconomic model for the future are outlined. These include stock flow consistency, an understanding of the significance of a fiat currency, the role of the central bank with respect to liquidity management and its inability to directly control the volume of bank lending, and a recognition that the choice of the exchange rate regime is crucial in determining the extent to which a government can exercise discretion in policymaking.

TOUR OF THE BOOK

Chapter Outline

24.1 Introduction
24.2 The Balance of Payments
24.3 Essential Concepts
24.4 Aggregate Demand and the External Sector Revisited
24.5 Trade in Goods and Services, Product Market Equilibrium and the Trade Balance
24.6 Capital Controls
Conclusion
References

Learning Objectives

- Understand the components of the balance of payments and their interrelationship.
- Acknowledge the distinction between the nominal and real exchange rates.
- Analyse the role of trade in the determination of equilibrium national income.

Chapter Outline

Each **Chapter Outline** signposts its top level subheadings on the first page to indicate what you will cover, while **Learning Objectives** are also listed to flag up the most important take home messages to consider.

Boxed Text

General **boxed text** includes digressions from, and extra discussion of, material mentioned in the main text.

BOX 7.1 RULES OF ALGEBRA

Addition and subtraction

In general, what we add to or subtract from one side of the equation, we have to add to or subtract from the other side to maintain the equality.

Given an equation $y = x$, then we know that the equivalent expression is $y \pm z = x \pm z$. So, for example, $y = x$ is equivalent to $y + 2 = x + 2$.

We can also substitute an expression from one equation into another and maintain equivalence. For example, we might have $y = 2a$, and $a = 6z$. In this case, it is equivalent to write $y = 2(6z) = 12z$.

Multiplication and division

Given an equation $y = x$, then we know that the equivalent expression is $3y = 3x$ or $y/3 = x/3$. If we multiply or divide the left-hand side of the equation by a variable (or more complex algebraic expression) then we have to multiply or divide the right-hand side of the equation by the same variable (or expression). Dividing by zero is not allowed, however, and multiplying by zero is not very helpful!

REMINDER BOX

The sustainable goal for a government should be to maintain full employment and price stability and allow its fiscal balance to adjust accordingly to ensure aggregate demand is consistent with those goals. A sovereign, currency-issuing government can always meet those goals if it chooses.

Reminder Boxes

Reminder Boxes indicate wherever new learning or discussion requires you to revisit ideas from previous chapters.

TRY IT YOURSELF

You might like to input the data into a spreadsheet and compute the present value of the revenue stream in Equation (25.13) using a discount rate of 15.1 per cent. You should verify that it is equal to $10,000, which is exactly the present value of the initial outlay. The actual result you get may not equal $10,000 exactly but this is due the approximate iterative solutions used by the spreadsheet.

Try It Yourself

Try It Yourself boxes invite you to take on worked examples or simple exercises to illustrate newly-covered material.

Conclusion, References and Endnotes

Conclusion, **References** and any **Endnotes** appear as shaded text at the end of each chapter.

PREFACE

This book presents a comprehensive, university level study course in Macroeconomics from a Modern Monetary Theory (MMT) perspective.

Our approach is grounded in the operations of real world institutions, and our approach clearly identifies the policymaking capacity of central governments. The pedagogy thus starts by putting the currency-issuing government at the forefront.

We want students to understand how a modern monetary system operates, how the government and non-government sectors interact, how the central bank and the banks interact, how the labour market works, how trade and capital flows impact on economic outcomes and much more.

Students will appreciate what the capacities of a currency-issuing government are and how fiscal and monetary policy can be used purposefully to enhance the well-being of the nation.

Unlike earlier pluralist approaches to macroeconomics in which the teaching sequence begins with an exposition of the standard mainstream macroeconomics (dominated by the New Keynesian approach) and only then qualifies that conceptual structure and framing with some 'real world' criticisms and quibbles, we feel students are better informed if we build the narrative from the ground up, based on an understanding of how the monetary system actually operates. In adopting that approach, we are informed by our view that the mainstream macroeconomic approach does not provide coherent knowledge upon which to understand those real world monetary operations.

We believe that students who are taught in the mainstream tradition are introduced to concepts and explanations of how the monetary system operates which are simply incorrect when applied to the real world. The consequences of these major flaws of reasoning have, at times, been devastating. Just think about the Global Financial Crisis (GFC). Mainstream economists did not see it coming, and, when the world was on the precipice of a total financial collapse, many advocated government spending cuts to maintain low fiscal deficits. Alternatively, think about the intractable problems within the Eurozone – problems that were identified by MMT economists even before the launch of the euro, but which still are beyond the grasp of mainstream economists.

Among other mistakes, the mainstream approach was used to form predictions that quantitative easing would lead to accelerating inflation. It was thought that the rising fiscal deficits would push up interest rates and bond yields, and that governments would run out of money. Also, that adoption of the euro would allow European nations to converge toward high and sustainable growth rates with rising living standards for everyone. None of those predictions turned out to be accurate, yet governments that listened to this sort of advice forced millions of workers to lose their jobs and ensured that economic recovery would be a long, drawn-out and painful process.

The literature shows that economists working in the MMT tradition were much more prescient in understanding why the GFC occurred and why it quickly led to an even worse crisis in the euro area, and what the likely outcomes of the subsequent government interventions would be.

We thus believe that an approach that starts with a sound understanding of how the monetary system works and the capacities of the government as the currency issuer within it – the hallmark of the MMT approach – provides a much sounder platform for students to learn about the economy.

While we present a detailed descriptive understanding of the day-to-day operations of economic institutions (government, the central bank, commercial banks, households, firms, trading entities, and so on), we are also mindful of the past. In order to provide students with a contextual perspective, we also introduce them to a rich historical analysis, both in terms of the history of economic thought itself, and the economic history of nations. There is much to learn from history. Many of the current debates about policy options are just repeats of debates of yesteryear. More pointedly, some of the resulting propositions being offered today were already categorically shown to be poor options in the past. We believe that students need to be aware of these issues, so as to be better informed of the viable policy options available to government.

We believe the concepts drawn from cognitive sciences such as framing and language are important components in conducting an informed economic narrative and in report writing. In that context, a highly innovative feature of the book is to show students how we use language to frame economic narratives, to convey economic concepts and form policy options. In this way, we place macroeconomics in a broader social science context, which we consider enriches the learning experience for students.

We have not shied away from recognising the benefits of formality: the use of mathematics and statistical techniques. The book provides a solid introduction to more technical aspects of the subject – students will gain knowledge of mathematical techniques, not for their own sake, but because they help us gain a more effective understanding of the material covered.

Further, students will learn how to deal with real world data and apply that knowledge to real world situations. We see that as being essential preparation for students once they enter the professional stage of their lives and are required to prepare meaningful economic commentary.

Macroeconomics is an exciting area of study not the least because it has a profound influence on our everyday lives and the prosperity of our nations.

We want students who use this book to gain a thorough knowledge of the field of macroeconomics in order to equip them with the skills necessary to make sense of the important macroeconomic debates and to be able to constructively participate in those debates.

We hope you enjoy the journey.

William F. Mitchell
L. Randall Wray
Martin J. Watts
October 2018

ACKNOWLEDGEMENTS

The overall project encapsulated in this book is dedicated to the millions of nameless unemployed workers who have been denied the opportunity to live fulfilled lives by the implementation of neoliberal economic policies abetted by orthodox economists who, unfortunately, bear none of the costs of their folly.

The authors would like to specially acknowledge the valuable administrative and editorial assistance they received from Melinda Hannan (at CofFEE).

The authors have appreciated great feedback from many students at the University of Missouri-Kansas City and at the University of Newcastle who have been exposed to this material as it has evolved.

They also would like to thank the input and guidance arising from countless discussions with colleagues, including Warren Mosler, Mat Forstater, Stephanie Kelton, Pavlina Tcherneva, Eric Tymoigne, Yeva Nersisyan, Flavia Dantas, James Juniper, Andrew Nadolny, Tim Sharpe, John Henry, Jan Kregel, Steve Fazzari, Wynne Godley, and Hyman Minsky.

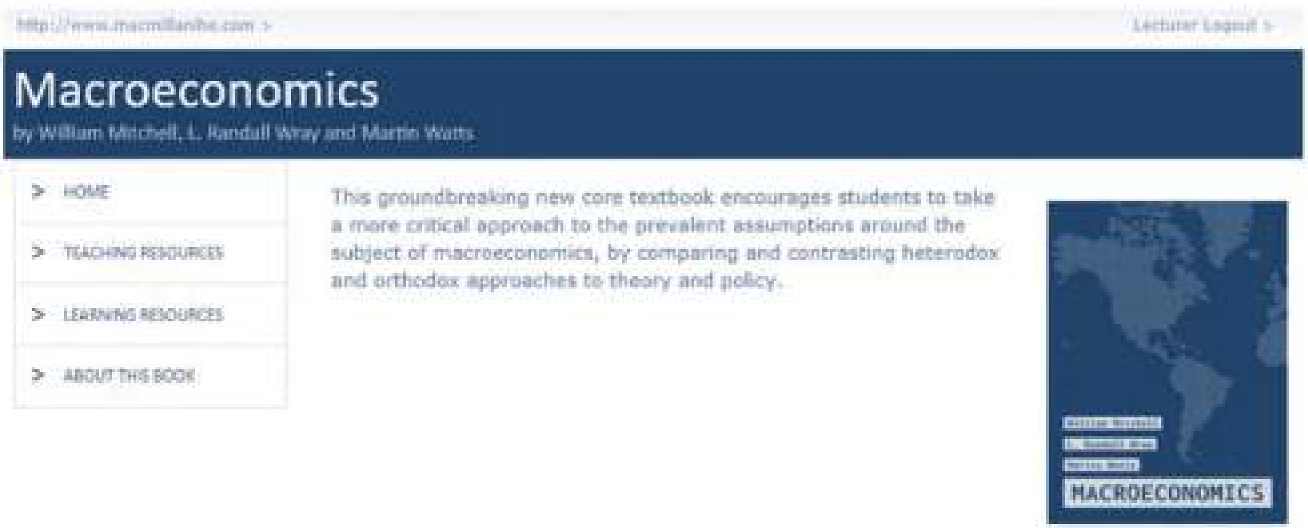
On a personal note, Bill would like to thank Louisa for her constant support, critical oversight and love. Randy would like to thank Shona, Xinhua, Shane, Alina, and Allison for their patience. Martin would like to acknowledge the forbearance of Linda, Rachel, Ben and Clare who expected the textbook to be finished some years ago! Also, Martin thanks niece Gemma who has maintained her interest in this project, despite her disciplinary areas of expertise being law and psychology!

While many have offered help and advice, all errors remain the responsibility of the authors.

WEBSITE MATERIALS

Companion Website

Visit the book's companion website at **www.macmillanihe.com/mitchell-macro** for extra resources to support and enhance your teaching and learning:



For Students:
- videos covering each major topic
- statistical data sets for recreating and experimenting with the graphs from the book
- additional further reading, weblinks and reference resources

For Lecturers:
- a Testbank of assessment questions
- a Lecturer Guide for using *Macroeconomics* in your teaching and module planning

PART A

INTRODUCTION AND MEASUREMENT

CHAPTER

1 INTRODUCTION

Chapter Outline

Learning Objectives

- Understand that macroeconomics analyses the behaviour of aggregates, such as employment, unemployment, GDP and inflation, whereas microeconomics studies the behaviour of individual economic agents, notably households and firms.
- Recognise that macroeconomics is a contested discipline with two broad schools of thought that differ in terms of their perspectives on the effectiveness of markets and the role of government.
- Acknowledge that social science disciplines and physical science disciplines each have their own language in the form of concepts and theories, and these provide the basis for understanding, and not merely describing, relevant phenomena.

1.1 What is Economics? Two Views

US President Harry Truman is said to have sought a one-armed economist because he was so frustrated by his economic advisors always telling him, "Well, on the one hand, we could do X, but on the other hand, we could instead do Y" – with 'Y' typically being the precise opposite policy to 'X'.

The story is of course funny, but it does highlight a problem that is ubiquitous to all social sciences. Unfortunately, economics is sometimes relegated to something like the 'study of business decision making', or even just a branch of mathematics, a view enabled in part by the heavy use of mathematics and models in much of the discipline. This view of economics as a 'decision science' assumes a highly artificial hypothesised world of hyper-rational automatons that consistently act to maximise pleasure and avoid pain. If this were true, of course it would have been easy for Truman's advisors to have come up with the one 'right' policy to follow.

This textbook will instead take a broader perspective of the discipline of economics, by including it within the social sciences. As Truman's experience illustrates, economics is as difficult as the other social sciences, such as psychology and political science, because it too concerns human behaviour, taking place in a sphere that we designate as 'the economy', which itself is hard to define and to delineate from other arenas of human interaction. Since the main topic of the social sciences – human behaviour – is complex, we often do not understand its causes, or even its nature, and much less do we know how to influence it in a desired manner. Even if we know the result we would like to achieve (say, smarter and happier kids), we do not know with certainty which policy choices would produce the desired outcome.

While we may think it is useful to separate 'the economy' from the rest of social life, and to apply 'economics' to the study of that area of life, we recognise that the division is necessarily arbitrary. In truth, there is no completely separate sphere of 'economic life', meaning that economics is linked to, and incorporates findings from, the other social science disciplines.

Further, we want to stress that there is no single 'right' way to do economics. In this textbook we will use a variety of methods and approaches to build our understanding of 'the economy'. We will occasionally bring in research and methods from other disciplines. We will use some mathematics and modelling. Because we believe that economic history, as well as the history of economic thought, helps us to understand today's economy, we will look back in time, both in terms of economic events and to examine the insights of the great thinkers of the past.

In the rest of this section we will briefly outline the two main approaches to economics taken by those thinkers, as well as by today's economists. It is always risky to pigeonhole individuals and their theories into categories. Just as a politician in any political party (say, the Labor Party in Australia, or the Republican Party in America) will hold many views that are shared by most members of that party, they will likely also hold some views that are more consistent with those of a rival party. This is true of economists too. Still, it is useful to identify the two broad approaches to economics that have dominated much of the debate over the past two centuries.

Recalling the story about President Truman's frustration, we can think of the 'two hands' of economics as the orthodox, or neoclassical approach, of which a number of related strands have emerged, and the heterodox or Keynesian/Institutionalist/Marxist approach, which also is a 'broad church'. Let us examine each in turn, while recognising that we must generalise.

Orthodox, neoclassical approach

In the neoclassical approach, an important assumption about human nature is made: individuals seek to maximise pleasure and avoid pain. Pleasure is defined as 'utility', so individuals pursue utility-maximising behaviour and avoid the 'disutility' of pain. Further, rational individuals are self-interested, seeking to maximise their own utility, and they do not receive either utility or disutility from the experiences of others. Neoclassical economics presumes individuals are 'rational', meaning that they maximise utility, given constraints. If there were no constraints, neoclassical thinking indicates that individuals would maximise utility to infinity. However, individuals are constrained by the resources that they have at their disposal, which are referred to in neoclassical economics as 'individual resource endowments'. Mutually beneficial exchange redistributes resources according to preferences, increasing the utility of both parties to the trade.

In the hypothesised free market, exchanges take place at competitively determined relative prices. (Relative prices are ratios; for example: one deer = three beavers = six rabbits = two bushels of wheat = ten hours of labour services.) Participants in markets take relative prices as signals. Relative scarcity will cause the price of a traded commodity to rise, inducing suppliers to produce more of it, and buyers to demand less. For example, if the supply of students trained in economics is insufficient to meet the demand for economists, the relative wage of economists to that of historians, say, rises. This signals to students that they ought to switch from the study of history to the study of economics. At the same time, employers try to find acceptably close – but cheaper – substitutes; say, political science students. As the supply of economists increases, the relative wage advantage for students trained in economics falls. Of course, other factors enter into decisions, but the important point is that relative prices function as signals to both suppliers (economics students) and demanders (employers of economists).

Equilibrium is defined as the set of relative prices that 'clear' markets; a 'general equilibrium' is a complete set of prices to clear all markets. Adam Smith's famous metaphor of an **invisible hand** is invoked to explain how markets are guided toward equilibrium prices. One interpretation of the 'invisible hand' analogy is that by producing market clearing prices, the market provides the signals that guide individuals to maximise their utility while also providing the social or public good of ensuring that demand and supply are equilibrated. To understand the invisible hand idea, we could envision a farmers' market held in the public square on a weekend. The farmers bring their fruits and vegetables early on Saturday morning, advertising the prices at which they are willing to make sales. Over the course of the day, some discover they've set them too low (facing a brisk pace of sales that would deplete their inventory too soon) while others have set them too high. At the same time, consumers adjust their reservation prices (the maximum they will pay for a given quantity/quality) as well as their desired quantities in light of offer prices. Prices are adjusted over the weekend to try to maximise revenue while ensuring the farmers do not have to cart home unsold produce. In this narrative, rational behaviour by producers and consumers will adjust prices so that 'supply equals demand'; that is, all fruits and vegetables have been sold.

The hand is 'invisible', guiding individuals and the economy as a whole toward equilibrium, with no need of an authority. For that reason, there is little need for the government to manage the economy. While it is a bit of an analytical leap, the next step in this 'free market' narrative is to stretch the market analogy to the economy as a whole. Surely if all prices and wages were flexible, every market, including the markets for labour of every kind of skill, would clear, with demand equalling supply in each of the markets? Would it be rational for any individual supplier or demander to stubbornly refuse the guidance provided by the invisible hand? It seems that the overall economy might reach a grand **general equilibrium** with a set of prices and wages (one for each type of product or input to production) that clears every market.

Certainly government has some role to play in setting and enforcing rules, in providing national security, and (perhaps) for providing a social safety net. But according to this interpretation of Smith, there is no need for the government to direct individuals to serve the public interest because by reacting to price signals and pursuing their own interests, individuals actually act in the public interest.

There is one more important conclusion to be reached by neoclassical economics: "you deserve what you get". If we all come to the free market to make mutually beneficial exchanges, we are all seeking to maximise our own individual utility subject to our resource constraints. The equilibrium allocation can be construed as being 'fair'. That does not mean that the allocation is equal. Some will have more (and achieve greater utility) and others will have less but that is because some start with greater endowments (of resources, ability, and drive).

Technically, the idea is that one receives an allocation of resources based on one's own contribution to the market. If your final allocation is low, it is because you did not bring enough to market: perhaps you were born with few resources, you made a constrained choice to obtain little education, and you prefer leisure over work. In other words, you have no one to blame for your meagre allocation but yourself.

To be sure, neoclassical economics also allows for bad luck, congenital disabilities, and so on. Hence, there is a role for social policy to get involved in altering the allocation in order to protect the poorest and least advantaged. However, generally speaking, allocations ought to be left to the market because it will reward each participant according to their productive contributions to the market; a dimension of fairness.

In recent years, the neoclassical approach to economics has been invoked in support of the conservative backlash against post-Second World War economic and social reforms in Western nations. (This movement is generally called 'neoliberal' outside the USA or 'neoconservative' within the USA.) This 'anti-government' stance is closely associated with the terms in office of President Ronald Reagan in the USA and Prime Minister Margaret Thatcher in the UK. When running for President in 1980, Reagan promised to "get the government off the backs of the people", while Thatcher was famous for arguing that there is no such thing as society, reflecting the individualistic framework shared by neoclassical economics.

Downsizing government, and especially reducing the social safety net, is consistent with the view that government only needs to 'get the incentives right', and then the 'free market' will maximise individual welfare while the invisible hand will ensure that signals coming from markets guide individuals to do what is best for the economy as a whole.

While neoliberal/neoconservative policies are most closely associated with conservative political parties, even the moderate and social democratic parties adopted these policies throughout the 1990s and 2000s. For example, US President Clinton (a Democrat) echoed President Reagan's distaste for social welfare programmes when he promised to "end welfare as we know it" in his 1992 election campaign. He eliminated the biggest anti-poverty programme (Aid to Families with Dependent Children) and replaced it with a term-limited programme that tried to force aid recipients to work for their benefits ('workfare' rather than 'welfare').

Outside the USA other left-wing parties such as the Labour Party in the UK pursued similar strategies (such as 'work for the dole'). Many European social democratic parties have overseen the fiscal austerity, privatisation and deregulation that have occurred in the Economic and Monetary Union (the 'Eurozone'). Neoclassical economic theory provided strong justification for these economic and social policy changes as politicians invoked the so-called benefits of a greater reliance on 'market outcomes' while reducing 'government interference' in the workings of the 'invisible hand'.

Since the 1980s, the neoclassical domination of the public debate has been so powerful that most political parties, at least in terms of economic policy, have fallen into step with these ideas. We shall see that many of the policy decisions that were based on an adherence to this school of thought have led to poor outcomes.

Finally, let us turn to the neoclassical definition of economics, as it provides a very nice summary of the approach taken.

Neoclassical Definition of Economics: the study of the allocation of scarce resources among unlimited wants.

This definition is often framed as 'the economic problem' – that is, while resources are scarce, our wants are unlimited. The 'problem' is that we cannot ever satisfy our wants. While we all try to 'maximise utility', resource constraints prevent us from ever achieving maximal bliss. For this reason, many call economics 'the dismal science' in recognition of the unsolvable nature of 'the problem'.

Another statement commonly attributed to economists is that "there's no such thing as a free lunch", which also derives from the neoclassical definition of economics. In other words, since resources are scarce, there is always a trade-off. If we move resources from one use to another, we necessarily reduce enjoyment of the first use in favour of enjoyment of the second. For example, if we want to have more 'guns', we must have less 'butter', and if we want to improve the standard of living enjoyed by 'Bob', we must reduce the living standard of 'Jill'.

Strictly speaking, this would be true only at full employment of all resources. However, with the invisible hand guiding the allocation of resources, flexible relative prices ensure that all scarce resources are fully employed. The idea is that prices will always fall until supply equals demand so that no resource is left idle.

Note also that the trade-off might only be temporary. For example, if we move resources out of the production of consumption goods and into the production of investment goods that raise productive capacity, then in the future we can have (produce) more consumer goods. Through economic growth we can increase the level of production so that both 'Bob' and 'Jill' can have more. This does not contradict the admonition that there is no free lunch, however. If we are to have more production in the future, we need to be willing to sacrifice some consumption today.

We will have much more to say about the neoclassical approach later in the text. However, it is time to move on to the alternative approach.

Heterodox approach – Keynesian/Institutionalist/Marxist

There is a second, important tradition in economics that adopts a quite different framework. Unfortunately, there is no strong consensus about what to call it. Sometimes it is called 'non-orthodox', which appears to define it in opposition to 'orthodox' or neoclassical economics. In recent years, many of those working in this tradition have settled on the term 'heterodox', but that adjective too is usually defined as 'not in agreement with accepted beliefs'. Yet from the end of the Second World War until the early 1970s, those views now labelled 'heterodox' economics were actually dominant, and it was the 'orthodox' views that could be considered 'unorthodox' in the sense that they were not the majority opinion!

Further, while all orthodox theorists substantially accept the tenets of neoclassical theory, 'heterodoxy' is made up of a number of well-established and coherent economic schools of thought.[1] While these share a common approach, they also deviate from one another in important ways. The three most important of these schools of thought are the Marxist (following the work of Karl Marx), the Institutionalist (following the work of Thorstein Veblen), and the Keynesian (followers of John Maynard Keynes).[2]

What are we to do? In spite of the semantic objections we have raised, we will conform to the convention and call this second approach the heterodox or Keynesian/Institutionalist/Marxist approach. Let us examine their shared framework.

First, according to this approach there is no such thing as 'natural' human behaviour; rather, it is shaped and changed by institutions, culture and society. There is nothing inherently 'natural' about self-interested (or better, 'selfish') behaviour, nor would such behaviour be 'rational' in the neoclassical sense. Humans are social animals and in many cultures, selfish behaviour is punished and selfish individuals are ostracised. Since human survival requires cooperation, selfishness would actually be irrational because losing the support and resources of the group would reduce one's chances of survival. In all known societies, elaborate rituals and traditions have evolved to promote cooperation and even sacrifice for the common good.

Human behaviour varies significantly across societies, and the economic system is one factor that helps to determine appropriate behaviour within any particular society. Self-interested behaviour is more acceptable in some societies than in others. It is not a coincidence that neoclassical economic theory was developed largely in Western capitalist societies, and particularly in the UK. The 'rational' behaviour attributed by neoclassical economists to all humans actually forms a reasonably accurate description of the behaviour of early British capitalists. In the social environment in which they operated, pursuit of their own self-interest without regard to the welfare of others (especially that of their employees) may have increased their probability of success as capitalists. Further, they operated in a hostile political climate in which the Crown and the land-owning aristocracy wanted to protect the dominance of an agrarian economy that favoured them, and to maintain or even increase their own share of the nation's rather feeble output. Government 'intervention' was almost always a bad thing from the perspective of the first capitalists because government operated substantially in the interest of the Crown and the aristocracy.

We will not go into economic history now. What we wish to emphasise is that human behaviour is surprisingly malleable and influenced in a complex manner by custom and tradition.

Furthermore, we cannot know for certain that any action we take is truly 'utility maximising'. Should I buy the Renault or the Mazda motor car? After the decision has been made I might have a better idea of the better choice with the passage of time, but it is more probable that, even a decade down the road, I will not know which would have been best. Obviously, that choice is relatively unimportant and simple compared to most economic choices one must make. In truth, we almost never know whether we made the right decision for 'maximising' utility, even with hindsight.

According to the heterodox approach, decisions and behaviour depend on a range of other factors, including uncertainty, power, discrimination, prejudice and segregation. The range of options that are realistically available to individuals depends on their status, social class, race, religion and gender, for example. These 'noneconomic' factors heavily influence and even constrain our choices.

Heterodox economists of all persuasions reject the notion that economic outcomes are arbitrated by an impersonal market that only seeks to equilibrate 'demand and supply'. In the real world, market prices are largely administered by firms with market power. Wages are set not to 'clear' the labour market, but rather to reflect the outcome of contested bargaining processes between workers and the representatives of capital. Capitalism is a system defined by class conflict. In general, workers want to earn as much as they can for the effort they expend, while bosses want workers to produce as much as they can while paying them as little as possible. And, as will be discussed later, unemployment cannot be eliminated through wage reductions that eliminate excess labour supply; indeed, wage reductions can reduce the demand for labour and thus increase unemployment. More generally, wages and other prices are not simply signals of the invisible hand, but rather determine incomes and thus influence business sales and decisions going forward. For that reason, price and wage determination are not usually left to the invisible hand of the market.

Heterodoxy holds a different view of the so-called 'economic problem' of scarce resources and unlimited wants. Wants are largely socially created, and there is nothing natural about humans having 'unlimited' wants. While it is true that modern advertising operates to continually expand our desires, this can be countered through education. Further, resources are also largely socially created. While it is true that some natural resources have a limited supply, innovations continually produce substitutes. For example, Western societies faced their first major energy crisis in the 19th century when whalers had significantly reduced the number of whales, the source of whale oil used for lighting and other purposes. However, the production of petroleum and then electricity quickly replaced the need for whale oil.

Moreover, the most important resource in any economy is labour. Ironically, in capitalist economies labour is virtually always in excess supply, that is, many workers are left unemployed. It is ironic that neoclassical economics starts from the presumption that resources are scarce, when the obvious empirical fact is that labour is underutilised. Any theory that begins with the presumption that labour is always fully employed, and hence scarce, is ignoring a glaring inconsistency.

Let us look at the heterodox definition of economics.

Heterodox Definition of Economics: the study of social creation and social distribution of society's resources.

Note that unlike the orthodox definition, heterodoxy focuses on the creation of resources. Further, most of that creation is a collective undertaking rather than an individual one: people work together to produce society's resources. Distribution too is socially determined, rather than being determined by a technical relation (one's contribution to the production process). For example, labour unions engage in collective bargaining with their employers, who themselves band together to keep wages low.

The political process is also important in determining distribution; not only does government directly provide income to (employ or support) large segments of society, but it also puts in place minimum wages, benefits, and working conditions that must be met by employers. Government is also a creator of resources; it is not just a user of them. It organises and funds innovative research and development (often in its own labs) that is then used to create resources (frequently by private firms). It also purchases directly from firms, encouraging them to increase hiring and output. Not only do these government activities increase production, but they also affect distribution. This is well understood by voters and their representatives in government because policy creates winners and losers, and not usually in a zero-sum manner. Thus some policies can create more winners while others might create more losers.

Power, discrimination, collusion and cooperation all play a role in determining who gets what. The point is that society does not have to let 'the market' decide that women should be paid less than men for example, or that those with less education should remain jobless and thus poor.

Economics, like all social sciences, is concerned with a society that is complex and continually undergoing change. Since economists study human behaviour in the economic sphere, their task is very difficult. Whatever humans do, they could have done something different. Humans have some degree of free will, and their behaviour is largely based on what they think they ought to do. That in turn depends on their expectations of an unknowable future. They do not know precisely what the outcome of their actions will be, and they do not know what others will do.

Indeed, humans do not know exactly what happened in the past, nor do they fully understand what is happening today. They must interpret the environment in which they live, and realise that they cannot fully understand it. They can never know if they have truly 'maximised' their pleasure. They make plans in conditions of existential uncertainty, and do the best they can, given their circumstances. Their actions are almost always taken with consideration given to the impacts on others. Humans are above all social animals and that is why economics must be a branch of the social sciences.

What do economists do?

Like sociologists and political scientists, economists are trying to understand particular aspects of human behaviour; for example decisions about levels and patterns of spending, choices about enrolment in post-school education and types of employment to pursue. We have argued above that all these are influenced by institutions, culture and society, at least as much as they are by 'purely economic' variables such as income, the prices of goods and prospective wage rates for different occupations. In microeconomics our focus is the behaviour of individual consumers and firms, whereas in macroeconomics the focus is the aggregate impacts of these decisions on outcomes at a national level, including total output and employment and the rate of inflation. We elaborate on these definitions of microeconomics and macroeconomics below.

In trying to understand particular forms of economic behaviour, we need to develop theories that require us to decide on those factors that we think influence particular economic decisions. In other words, we need to make simplifying assumptions, which means we necessarily ignore those factors that we consider to be irrelevant. Otherwise we would be trying to replicate the complex reality as we see it; engaging in description rather than theorising.

In the development of theory, we formulate concepts that can be viewed as the building blocks of theory. A model can be seen as the formalisation of a theory (see below). To understand any theory (model), it is important that students comprehend its underlying concepts.

Social scientists seek to test their abstract theoretical models, expressed in the form of conjectures about real world behaviour, using the empirical data that the real world provides. For example, we might form the conjecture that if disposable income (that is, income remaining after the payment of taxes) rises, household consumption will rise. We would then collect the relevant data for disposable income and household consumption and any other information we thought might bear on the relationship and use various statistical tools (for example, regression analysis) to enumerate the relationship between disposable income and household consumption to see whether our conjecture was data consistent.

In engaging in this sort of exercise, the responsible social scientist is not seeking to establish whether the theoretical model is true, because that is an impossible task given there is no way of knowing what the truth is anyway. Instead, we seek to develop theories or conjectures that provide the best correspondence with the empirical world in which we live. This means that our current, accepted body of knowledge comprises theories and conjectures that explain the real world data in the most comprehensive way possible compared to the competing theories.

Further, we can rarely completely refute a theory. As President Truman complained, there are two or more sides to the most important economic questions, so there are competing theoretical approaches yielding different conclusions. Even when a researcher utilises the analysis of relevant data (which often entails the use of econometrics), they can never refute a theory with 100 per cent confidence. Often the acceptance of a theory is driven by ideology and politics, rather than a balanced assessment of the competing theories and associated evidence.

Implications for research and policy

Many students, like President Truman, find the inability of economists to come up with definitive answers to economic questions to be rather frustrating. Here it is important to emphasise that, like physical sciences and other social sciences, economics is a contested discipline, as illustrated by our brief discussion of the two schools of thought above. Students will be exposed to some other major contemporary debates in macroeconomics in Chapter 31.

If there are long-standing debates in economics (and other disciplines) which appear to be unresolved, how can there be progress in our understanding of economic phenomena? This is an important question because decisions made by macroeconomic policymakers have profound effects on the welfare of the population in terms of, for example, employment opportunities and wages.

1.2 Economics and the Public Purpose

The households and business firms in a modern capitalist economy make many of the important economic decisions that contribute to determination of the level of employment and output, the prices and composition of that output and the distribution of income. Claims are sometimes made that a 'free market' economy comprised of individuals seeking only their own self-interest can operate 'harmoniously' as if guided by an 'invisible hand' (see Section 1.1). In fact, economists had rigorously demonstrated by the 1950s that the conditions under which such a stylised economy could reach such a result could not exist in the real world. In other words, there is no scientific basis for the claim that 'free markets' are best.

In any case, these claims – even if true for some hypothesised economy – are irrelevant for the modern capitalist economies that actually exist. This is because all modern capitalist economies are 'mixed', comprising huge corporations (including multinational firms), labour organisations and a significant government relative to the size of the economy. Individuals and firms operate within socio-political and cultural economic structures that are constraining but also enabling.

Sometimes the goals of individuals and firms coincide with what might be called 'the public purpose', while often they do not. In this section we will discuss the public purpose and the role played by government in trying to align private interests with socially progressive goals.

What is the public purpose? It is not easy to define or to identify. One of the basic functions of any social organisation is to provide the necessary food, clothing, shelter, education, health care, legal framework and socialisation for the survival of members of the society.

While the subject of this course is economics, there is no sharp distinction between the sphere of economics and the spheres of other social sciences that study social processes. We usually think of the economy as that part of the social organisation that is responsible for the provision of the material means of survival: food, clothing, shelter and so on. However, the economy is always embedded in the social organisation as a whole, affecting and affected by culture, politics and social institutions.

Even if we can agree that any successful economic organisation should be able to produce adequate food for its population, that still leaves open many questions: What kind of food?; How should it be produced?; How should it be distributed?; and even What does 'adequate' mean?

Further, no society comprises harmonious individuals and groups. There are always conflicting claims and goals that must be moderated. There is no single, obvious public purpose to which all members of a society wish to strive. Even if we can identify a set of goals that the majority of society would like to work toward, that set will surely change over time as hopes and dreams evolve. The public purpose is an evolving concept.

The position taken in this book is that there is no 'invisible hand' that ensures that private interests are consistent with the public purpose. Indeed, the economy is just one component of the social organisation that is necessary to establish the always evolving public purpose and to work towards its achievement.

The 'market' is just one institution among a wide variety of social organisations working to delineate social goals that comprise its social and private purposes. Other institutions include political organisations, labour unions, manufacturers and NGOs (non-governmental organisations).

As we noted at the beginning of this chapter, the national government plays an important role in society because it can help to identify the social purpose and to establish a social structure in which individuals and groups will work toward achieving the public purpose.

While it is admittedly difficult to outline what defines the public purpose, it is possible to identify widely accepted goals. For example, the United Nations Universal Declaration of Human Rights (1948) commits signatory nations to a common set of relatively well-defined goals.

The declaration is outlined on the United Nations Home Page (United Nations, 1948):

> *Now, Therefore THE GENERAL ASSEMBLY proclaims THIS UNIVERSAL DECLARATION OF HUMAN RIGHTS as a common standard of achievement for all peoples and all nations, to the end that every individual and every organ of society, keeping this Declaration constantly in mind, shall strive by teaching and education to promote respect for these rights and freedoms and by progressive measures, national and international, to*

secure their universal and effective recognition and observance, both among the peoples of Member States themselves and among the peoples of territories under their jurisdiction.

The Articles that define the Declaration include:

- Everyone has the right to life, liberty and security of person.
- No one shall be held in slavery or servitude; slavery and the slave trade shall be prohibited in all their forms.
- Everyone has the right to an effective remedy by the competent national tribunals for acts violating the fundamental rights granted them by the constitution or by law.
- Everyone has the right to freedom of movement and residence within the borders of each state.
- Everyone has the right to a nationality.
- Men and women of full age, without any limitation due to race, nationality or religion, have the right to marry and to found a family. They are entitled to equal rights as to marriage, during marriage and at its dissolution.
- Everyone has the right to own property alone as well as in association with others.
- Everyone has the right to freedom of thought, conscience and religion; this right includes freedom to change their religion or belief, and freedom, either alone or in community with others and in public or private, to manifest their religion or belief in teaching, practice, worship and observance.
- Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.
- Everyone has the right to freedom of peaceful assembly and association.
- Everyone has the right to take part in the government of their country, directly or through freely chosen representatives.
- Everyone has the right of equal access to public service in their country.
- Everyone, as a member of society, has the right to social security and is entitled to realisation, through national effort and international co-operation and in accordance with the organisation and resources of each state, of the economic, social and cultural rights indispensable for their dignity and the free development of their personality.
- Everyone has the right to work, to free choice of employment, to just and favourable conditions of work and to protection against unemployment.
- Everyone, without any discrimination, has the right to equal pay for equal work.
- Everyone who works has the right to just and favourable remuneration ensuring for themselves and their family an existence worthy of human dignity, and supplemented, if necessary, by other means of social protection.
- Everyone has the right to form and to join trade unions for the protection of their interests.
- Everyone has the right to rest and leisure, including reasonable limitation of working hours and periodic holidays with pay.
- Everyone has the right to a standard of living adequate for the health and well-being of themselves and of their family, including food, clothing, housing and medical care and necessary social services, and the right to security in the event of unemployment, sickness, disability, widowhood, old age or other lack of livelihood in circumstances beyond their control.
- Everyone has the right to education. Education shall be free, at least in the elementary and fundamental stages. Elementary education shall be compulsory. Technical and professional education shall be made generally available and higher education shall be equally accessible to all on the basis of merit.
- Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits.

It is obvious that many of these human rights, especially near to the end of this list, are connected to the operation of the economy. For example, we argued above that any successful economy should provide adequate food, clothing and shelter, and many of the human rights listed in the UN Charter address the material well-being of a nation's population.

Further, other human rights that superficially appear to be unrelated to economic performance actually presuppose fulfilment of other human rights that are themselves directly related to material well-being.

For example, in a modern capitalist economy access to employment (one of the recognised rights) is necessary for full participation in society. Not only does a job provide income that allows one to purchase food, clothing, and shelter, but it also provides access to social networks, generates feelings of self-worth (since one contributes to social production), enhances social prestige and helps to provide for retirement in old age.

Indeed, employment has been shown to have a wide range of other benefits to individuals and to society including better physical and psychological health, reduced crime and drug abuse, lower child and spouse abuse, and greater participation in other social and political activities.

To be sure, the above list (which is itself only a partial listing of the agreed universal rights) includes many rights that have not been fully achieved even in the wealthiest and most democratic nations. In that sense, these rights are 'aspirational', with the signatory nations committing to striving toward achieving them. Again, if we look at the example of the right to work and to an adequate standard of living, those are rights that are routinely violated even in the best of times in the wealthiest of nations. Still, these universally recognised rights provide a measure against which nations can assess their progress.

We conclude with three important points.

First, the public purpose is broad and evolving over time, and for these reasons it varies across time and place. It should include rising living standards, particularly for those on low incomes. Environmental sustainability must be included. The reduction of racial, ethnic, and gender inequalities is an important component of public purpose. This must go beyond simple economic measures such as family income to include full participation in the life of the community. The public purpose also should include reductions of crime, corruption, cronyism, invidious distinction, conspicuous consumption and other social pathologies.

Second, the UN Charter lays out what it sees as 'universal' human rights. This is a useful, but not wholly satisfactory list to be included in a statement of the public purpose. What is considered to be a human right today would have appeared to be radically utopian a century ago; and the above list will no doubt appear far too cautiously conservative at some date in the future.

The public purpose is inherently a progressive agenda that strives to continually improve the material, social, physical, cultural, and psychological well-being of all members of society. It is inherently 'aspirational' in the sense that there is no end because its frontiers will continually expand.

Third, national governments as well as international organisations (such as the United Nations) must play an important role in shaping our vision regarding the types of societies to which we aspire. And beyond setting these goals, governments at all levels must take the lead in developing sets of institutions, rules of behaviour and sanctions for undesirable behaviour in order to move societies toward the achievement of these goals.

As an example, in the 1950s national governments and international organisations started to eliminate the devastating disease known as smallpox. While markets and for-profit production played a role in helping to develop vaccines, in distributing the vaccines, and in formulating information campaigns, private initiative alone would never have eliminated the disease.

The task was too big, it was not completely consistent with self-interested profit-seeking behaviour, and it required international cooperation beyond the reach of even the largest firms.

Hence, governmental organisations had to play a role.

With respect to the aspirational nature of the public purpose, successful elimination of smallpox would not be the end, but rather would serve as the beginning of a new campaign, to eliminate another disease, and then another and yet another.

Perhaps in a long-distant future, a human right to a disease-free life will be recognised, adding to an ever-increasing list of established rights that all nations would be expected to protect.

While we cannot, of course, imagine such a future, it was not so long ago that the US Congress did not recognise the voting rights of women and African Americans. Today, any nation that denies the vote to members of society on the basis of gender, religion, race or ethnicity, or national origin is considered to be in violation of human rights, and thus, to be an international pariah, even though such restrictions were considered acceptable just a few generations ago. For example, white US women over the age of 21 did not secure the vote until the 1920

Presidential election, while in the UK suffrage was extended to all women over the age of 21 in 1928. In Australia Aborigines were granted the right to enrol and vote in federal elections in 1962. Many developed countries that are seen as liberal democracies today did not give women or minorities the vote until well into the 20th century (1971 in the case of Switzerland for instance).

The public purpose is inherently progressive; it can never be finished.

1.3 What is Macroeconomics?

In macroeconomics we study the aggregate outcomes of economic behaviour. The word 'macro' is derived from the Greek word '*makro*', which means large, as we take an economy-wide perspective.

Thus, macroeconomics is not concerned with analysing how each individual person, household or business firm behaves; that is the domain of the other major branch of economic analysis, *micro*economics. Macroeconomics focuses on a few outcomes at the aggregate level and is considered to be the study of employment, output and inflation in an international context. A coherent macroeconomic theory will provide consistent insights into how each of these aggregates is determined and why it changes.

In this regard, there are some key macroeconomic questions that we seek to explore:

1. What factors determine the flow of total output produced in the economy over a given period and its growth over time?
2. What factors determine total employment and why does mass unemployment occur?
3. What factors determine the evolution of prices in the economy (inflation)?
4. How does the domestic economy interact with the rest of the world and what are the implications of that interaction?

A central idea in both macroeconomics and microeconomics is efficiency: getting the best out of what is available. The concept is extremely loaded and is the focus of many disputes, some more arcane than others. However, there is a consensus among economists that at the macroeconomic level, the 'efficiency frontier' (which defines the best outcome achievable from an array of possible outcomes) is normally summarised in terms of full employment. The hot debate that has occupied economists for years is the exact meaning of the term 'full employment'. We will consider that issue in full in Chapters 17 and 18.

But definitional disputes aside, it is a fact that the concept of full employment is a central focus of macroeconomic theory. Using the available macroeconomic resources, including labour, to their limits is a key goal of macroeconomics. The debate is over what the actual limit is. The related macroeconomic challenge is how to maintain full employment but at the same time achieve price stability, that is, the growth of prices at a low and stable rate.

The clear point is that if you achieve full employment and price stability then you will be contributing to the prosperity and welfare of the population by ensuring real output levels are high within an environment of stable prices.

This book develops a framework for understanding the key determinants of these aggregate outcomes – the level and growth in output, the rate of unemployment, and the rate of inflation – within the context of what we call a monetary system. All economies use currencies as a way to facilitate transactions. The arrangements by which the currency enters the economy and the role that the currency issuer, the national government, has in influencing the outcomes at the aggregate level are crucial parts of macroeconomics. Modern Monetary Theory (MMT), which is briefly outlined below, develops a macroeconomic framework that incorporates the unique features of the monetary system.

The macro model

To organise our thinking about macroeconomic relationships, we use a conceptual structure sometimes referred to in the economics literature as a model; in this case a **macroeconomic model**. A model is just an organising framework and represents a simplification of the system that is being investigated. In this textbook, we will develop a macroeconomic model that combines narrative and some algebra to advance your understanding of how the real world economy operates. We will necessarily simplify where complexity hinders clarity, but we will always focus on the real world rather than an assumed abstraction that has no relevance to the actual economy.

All disciplines develop their own language as a way of communicating. One might think that this just makes it harder to understand the ideas and we have sympathy for that view. But we also understand how useful it will be for students of a specific discipline, in this case macroeconomics, to be somewhat conversant with the language of the discipline they are studying.

In Chapter 7 – *Methods, Tools and Techniques* – we present the essential analytical techniques and terminology that are used to specify and solve macroeconomic models throughout this book. These tools and techniques are also deployed in the practical exercises that accompany this text and are to be found on the internet home page for the book (www.macmillanihe.com/mitchell-macro). Chapter 7 should be consulted regularly.

A macroeconomic model draws on concepts and algebraic techniques to advance our understanding of the main economic aggregates (such as output, employment and price level). This textbook design is unique because it specifically develops the Modern Monetary Theory (MMT) macroeconomic model, which will inform economic policy debates. We introduce that approach in the next subsection.

The MMT approach to macroeconomics

Modern Monetary Theory (MMT) is distinguished from other approaches to macroeconomics because it places the monetary arrangements at the centre of the analysis. As we will see, MMT builds on the insights of many economists who have worked in the heterodox tradition. It therefore rejects the main precepts of the orthodox neoclassical approach to macroeconomics. However, because it places an emphasis on monetary arrangements within the capitalist economy, it adds new insights that were not previously available within the heterodox tradition. Learning macroeconomics from an MMT perspective requires you to understand how money 'works' in the modern economy and to develop a conceptual structure for analysing the economy as it actually exists.

By placing government, as the currency issuer, at the centre of the monetary system, the MMT approach immediately focuses on how a government spends, and how that spending influences those aforementioned macroeconomic aggregates that we seek to explain. The framework will first provide a general analysis of government spending that applies to all currency exchange rate systems, before explaining the constraints (policy options) that apply to governments as we move from a flexible exchange rate to a fixed exchange rate system. We will consider how the design of the monetary system impacts on the domestic policy choices open to government and the outcomes of specific policy choices in terms of output, employment and inflation.

The most important conclusion reached by MMT is that the issuer of a currency faces no financial constraints. Put simply, a country that issues its own currency can never run out and can never become insolvent in its own currency. It can make all payments as they come due. For this reason, it makes no sense to compare a sovereign government's finances with those of a household or a firm.

Households and firms are users of the currency; they must obtain the currency in order to make payments as they come due. They must either earn income, or borrow, or sell assets to obtain the currency. They can be forced to default. But the sovereign currency issuer can never run out of its own currency. In later chapters, we will explain how sovereign currency issuers spend and why they can always afford anything that is for sale if it is priced in their own currency.

There is a caveat, however. Even a sovereign currency issuer can tie its own hands. This occurs if the government promises to deliver precious metal (such as gold) or a foreign currency in payment. It is not uncommon for governments to issue debt denominated in foreign currency. This is especially true of developing nations. In this case, they must obtain the foreign currency to service their debts. In the past, many governments promised to exchange their own currencies for gold or silver so, again, they had to obtain the gold or silver to meet these promises. Thus while these governments cannot run out of their *own* currencies, they can certainly run out of precious metal or foreign currencies and then be forced to default *on their promise to make payments in precious metal or foreign currency.*

Many people are unaware that a major historical event occurred in 1971, when US President Nixon abandoned gold convertibility and ended the system of fixed exchange rates that had existed in the Bretton Woods international monetary system since the Second World War. Under that system (as well as under the gold standard system that had existed since the late 19th century, with breaks for both world wars), currencies were convertible into gold, and exchange rates were fixed to the US dollar. As such, they had to operate their economies in such a manner as to accumulate gold or dollars. This usually meant adopting contractionary fiscal policy as well as maintaining high interest rates to ensure trade surpluses and strong currencies. However, after 1971, most governments floated their currencies and traded them freely on foreign currency markets. Occasionally, central banks would conduct what became known as a 'managed' float where they tried to limit the amplitude of movements that the free float would generate.

It is thus essential to understand the notion of a currency regime, which can range through a continuum from a fixed exchange rate system to a floating exchange rate system with varying degrees of exchange rate management in between. Understanding the way the exchange rate is set is important because it allows us to appreciate the various policy options that a currency-issuing government has in relation to influencing the main objects of our study; employment, output and inflation. It also allows us to deepen our understanding of the policy options available to a government which chooses to use a foreign currency, such as the Member States of the Economic and Monetary Union (the 'Eurozone').

A flexible exchange rate releases monetary policy from defending a fixed parity (exchange rate) against a foreign currency. Fiscal and monetary policy can then concentrate on ensuring domestic spending is sufficient to maintain high levels of employment. A consequence of this is that governments that issue their own currencies need no longer accumulate large reserves of foreign currencies to defend their exchange rates. The reality is that currency-issuing governments such as those of Australia, Britain, Japan and the USA can never run out of money. These governments always have the capacity to spend in their own currencies.

However, most of the analysis appearing in macroeconomics textbooks, which filters into the public debate and underpins the cult of austerity, is derived from 'gold standard' logic and does not apply to modern fiat monetary systems. The economic policy ideas that dominate the current debate are artefacts from the old system, which was abandoned in 1971.

One of the most basic propositions in macroeconomics that MMT emphasises is the notion that at the aggregate level, total spending equals total income and total output. In turn, total employment is related to the total output in the economy. So to understand employment and output determination we need to understand what drives total spending and how that generates income, output and the demand for labour.

In this context, we will consider the behaviour and interactions of the two economic sectors: government and non-government. Then we will unpack the non-government sector into its component sectors: the private domestic sector (consumption and investment) and the external sector (trade and capital flows). In Chapter 4 we analyse in detail the so-called National Accounts, drawing on these broad macroeconomic sectors. This approach is called the *sectoral balance approach*, and builds on the accounting rule that a government deficit (or surplus) must be exactly offset by a surplus (or deficit) in the non-government sector. The non-government sector comprises the private domestic and external sectors. So a more general observation is that the sum of the sectoral balances nets to zero when we consider the government, private domestic and external sectors.

If one sector spends more than its income, at least one of the others must spend less than its income because for the economy as a whole, total spending must equal total receipts or income. While there is no reason why any one sector has to run a balance between its spending and income, the National Accounts framework shows that the system as a whole must. Often though, but not always, the private domestic sector runs a surplus (spending less than its income). This is how it accumulates net financial wealth. Overall private domestic sector saving (or surplus) is a leakage from the overall expenditure cycle that must be matched by an injection of spending from another sector. The current account deficit (the external sector account) is another leakage that drains domestic demand. A current account deficit occurs when the domestic economy is spending more overseas than foreigners are spending in the domestic economy. These concepts are developed in full in Chapter 6.

Here it is useful to differentiate between a stock and a flow. The latter is a magnitude per period of time. For example, spending is always a flow of currency per period (for example, households might have spent $100 billion dollars in the first three months of 2018). On the other hand, a stock is measured at a point in time. For example, a student's financial wealth could have consisted of a deposit account at a local bank, with a balance of $1,000 on 1 January 2018. We explain stocks and flows in more detail in Chapters 4 and 6.

The sectoral balances framework shows that a sectoral deficit (a flow per year, say) accumulates as a matter of accounting to a financial debt (a stock). On the other hand, a sequence of sectoral surpluses accumulates to a financial asset, which is also a stock. MMT is thus based on what is known as a stock-flow consistent approach to macroeconomics by which all flows and resulting stocks are accounted for in an exhaustive fashion. The failure to adhere to a stock-flow consistent approach can lead to erroneous analytical conclusions and poor policy design.

From the perspective of fiscal policy choices, an important aspect of the stock-flow consistent approach that will be explained in greater detail in Chapter 6 is that one sector's spending flow equals its income flow plus changes to its financial balance (stock of assets).

This textbook will show that a country can only run a current account deficit if the rest of the world wishes to accumulate financial claims on the nation (financial debt). The MMT framework also shows that for most governments, there is no default risk on government debt, and therefore such a situation is 'sustainable' and should not necessarily be interpreted to be undesirable. Any assessment of the fiscal position of a nation must be taken in the light of the usefulness of the government's spending programme in achieving its national socio-economic goals. This is what Abba Lerner (1943) called the **functional finance** approach. Rather than adopting some desired fiscal outcome (relationship between spending and taxation revenue), a government ought to spend and tax with a view to achieving 'functionally' defined outcomes, such as full employment.

On matters of terminology, we avoid using the term 'budget' to describe the spending and taxation outcomes for the currency-issuing government. Instead, we use the term **fiscal balance**. A government fiscal deficit occurs when its spending exceeds its taxation revenue, whereas a fiscal surplus occurs when government spending is less than its taxation revenue.

The use of the term 'budget' to describe the fiscal balance invokes the idea that the currency-issuing government faces the same financial constraints as a household when it is forming its budget. A careful understanding of the monetary system will make it obvious that the government is not a 'big household'. The government can consistently spend more than its revenue because it creates the currency. Households use the currency issued by the government and must finance their spending. Their access is constrained by the sources of available funds, including income from all sources, asset sales, and borrowings from external parties. Whereas households have to save (spend less than they earn) to spend more in the future, governments can purchase whatever they like, as long as there are goods and services for sale in the currency they issue.

A sovereign government must spend before it can subsequently tax or borrow. A household cannot spend more than its revenue indefinitely because continuously increasing private debt is unsustainable. The budget choices facing a household are thus limited and prevent permanent deficits. A currency-issuing government can never be revenue constrained in a technical sense and can sustain deficits indefinitely without solvency risk. In other words, our own personal budget experience generates no knowledge that is relevant to the consideration of government matters. This alternative narrative, which we present in this book, highlights the special characteristics of the government's currency monopoly.

Fiscal surpluses, which arise when a government's spending is less than the amount it takes out of the economy by way of taxation, do not provide governments with a greater capacity to meet future needs, nor do fiscal deficits erode that capacity. Governments always have the capacity to spend in their own currencies.

In summary, budget surpluses force the non-government sector into deficit and the domestic private sector is forced to accumulate ever-increasing levels of indebtedness to maintain its expenditure. We will explain why this is an unsustainable growth strategy and how eventually the private domestic sector is forced to reduce its risky debt levels by saving more. The resulting drop in non-government spending will reinforce the negative impact of the government fiscal surplus on total spending.

Fiscal and monetary policy

The two main policy tools that influence what is termed the demand, or spending, side of the economy are monetary and fiscal policy.

Fiscal policy is represented by the spending and taxation choices made by the government (the 'treasury'). The net financial accounting outcomes of these decisions are summarised periodically by announcements of the government's fiscal position. Fiscal policy is one of the major means by which the government seeks to influence overall spending in the economy and achieve its economic and social objectives.

This textbook will show that a nation will have maximum fiscal space (that is, capacity for government to use its fiscal policy tools of spending and taxation):

- If it operates with a sovereign currency; that is, a currency that is issued by the sovereign government and whose value is not pegged to foreign currencies; and
- If it avoids incurring debt in foreign currencies, or guaranteeing the foreign currency debt of domestic entities (firms, households, and state, province, or city debts).

Under these conditions, the national government can always afford to purchase anything that is available for sale in its own currency. This means that if there are unemployed resources, the government can always put them to productive use through the use of fiscal policy. To put it as simply as possible, this means that if there are unemployed workers who are willing to work, a sovereign government can afford to hire them to perform useful work in the public interest. As we have noted, from a macroeconomic efficiency perspective, a primary aim of public policy is to fully utilise available resources. Under these optimal conditions, the government is not revenue constrained, which means it does not face the financing constraints that a private household or firm faces in framing its expenditure decision.

The central bank in the economy is responsible for the conduct of monetary policy, which typically involves the setting of a short-term policy target interest rate. Since the 2008 global economic crisis the ambit of monetary policy has broadened considerably and these developments will be considered in Chapter 23.

The typical roles of a central bank include not only the conduct of monetary policy via the overnight interbank lending rate, but also operating the interbank clearing mechanism (so that bank cheques clear among banks), acting as lender of last resort (to stop bank runs), and regulating and supervising the banks.

MMT considers the treasury and central bank functions to be part of what is termed the **consolidated government sector**. In many textbooks, students are told that the central bank is independent from government. The MMT macroeconomic model will demonstrate how it is impossible for the central bank to work independently of the treasury if the monetary system is to operate smoothly.

Policy implications of MMT for sovereign nations

MMT provides a broad theoretical framework based on the recognition that sovereign currency systems are in fact public monopolies *per se*, and that the imposition of taxes coupled with insufficient government spending generates unemployment.

An understanding of this point will be developed to allow the student to appreciate the role that government can play in maintaining its near-universal dual mandates of price stability and full employment. The student will learn that two broad approaches to controlling inflation are available to government in designing its fiscal policy.

Both approaches draw on the concept of a **buffer stock** to control prices. We will examine the differences between the use of:

a. *Unemployment buffer stocks*: The neoclassical approach, which describes the current policy orthodoxy, seeks to control inflation through the use of high interest rates (tight monetary policy) and restrictive fiscal policy (austerity), and leads to a buffer stock of unemployment. In Chapters 17 and 18, students will learn that this approach is very costly and provides an unreliable target for policy makers to pursue as a means for inflation control; and
b. *Employment buffer stocks*: Under this approach the government exploits its fiscal capacity, which is inherent in its currency-issuing status, to create an employment buffer stock. In MMT, this is called the **Job Guarantee (JG)** approach to full employment and price stability. This model, which is considered by MMT to be the superior buffer stock option, is explained in detail in Chapter 19.

The MMT macroeconomic framework shows that a superior use of the labour slack necessary to achieve price stability is to implement an employment programme for those who are otherwise unemployed, which both anchors the general price level to the price of employed labour of this (currently unemployed) buffer and can produce useful output with positive supply side effects.

Conclusion

This chapter has emphasized that economics is a social science that studies "economic life". We defined the economy as that part of the social organization that is responsible for the provision of the material means of survival: food, clothing, shelter, and so on. However, the economy is always embedded in the social organization as a whole, affecting and affected by its culture.

There is no single 'right' way to do economics. Economic theories, as well as economists themselves, can be grouped into two main approaches: the orthodox, Neoclassical approach and the heterodox, Keynesian/Institutionalist/Marxist approach. We will see that public policy recommendations that follow from each of these two approaches are quite different because they are based on very different views as to how the economy works.

The heterodox approach, for example, envisions a more important role for the government to play in ensuring that the economy furthers the public purpose. This chapter discussed widely shared goals of public policy as enumerated in the United Nations Universal Declaration of Human Rights, many of which are related to economic issues such as the right to work.

Finally, the chapter defined the scope of macroeconomics as the study of the aggregate outcomes of economic behaviour. The Modern Money Theory (MMT) approach was introduced and distinguished from other approaches to macroeconomics because it places the monetary arrangements at the centre of the analysis. In particular, **MMT puts emphasis on the sovereign nature of currency and the policy implications that derive from the ability of a national government to issue its own currency**. This is a theme that will be examined throughout this text.

References

Lerner, A. (1943) "Functional Finance and the Federal Debt", *Social Research*, 10 (1), 38–51.

Samuelson, P. (1947) *Foundations of Economic Analysis*, Cambridge, MA: Harvard University Press.

United Nations (1948) United Nations Declaration of Human Rights, United Nations General Assembly, 10 December 1948, available at: http://www.un.org/en/universal-declaration-human-rights/ accessed 15 January 2016.

Endnotes

1. Note that the approach taken in this text, Modern Money Theory, falls within the heterodox camp. Indeed, it rests upon the foundations of many of the heterodox traditions.
2. A caveat is necessary here. Many of those who call themselves 'Keynesian', as well as the approach that is often presented in economics textbooks as 'Keynesian theory', are not heterodox. They are much closer to the neoclassical approach. Indeed, one of the founders of orthodox macroeconomic theory, Paul Samuelson (1947), called it the 'Neoclassical Synthesis' to indicate that while its foundations are neoclassical, some of Keynes' ideas are 'synthesised' or grafted onto that base. Heterodox followers of Keynes argue that such integration is not possible. We will revisit these issues in more detail in Chapter 27.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

2 HOW TO THINK AND DO MACROECONOMICS

Chapter Outline

Learning Objectives

- Recognise the importance of the fallacy of composition in understanding macroeconomics.
- Gain an awareness that macroeconomics is a highly contested discipline in terms of theory and policy prescription.
- Note the importance of referring to the stylised facts in analysing theory and policy prescription.
- Develop critical thinking skills about the working of a macroeconomy with its own sovereign currency.

2.1 Introduction

In Chapter 1, we noted that any science, whether physical or social, develops theories to gain an understanding of the specific phenomena that it is trying to explain. This necessitates abstraction.

In economics there are two broad schools of thought, which means that economics is a contested discipline, with ongoing debates about both theory and policy. In Chapter 1, we outlined the subject matter of macroeconomics and highlighted the distinct features of Modern Monetary Theory (MMT). Finally, we provided a discussion of macroeconomic policy objectives, by introducing the concept of public purpose.

All disciplines have their own language and way of thinking. In the next section, we argue that thinking as a macroeconomist is particularly challenging because the discipline is highly contested, with self-styled experts offering diverse views. An important contemporary example is the MMT rejection of the neoclassical claim that a currency-issuing national government is like a household and subject to the same type of 'budget' constraint. More generally some propositions, which make sense at an intuitive, personal level, fail to hold true at an aggregate level. This is referred to as the **fallacy of composition**.

A number of examples, both economic and non-economic, are provided. We then discuss what macroeconomics should be able to explain and outline two empirical examples relating to unemployment and the conduct of fiscal policy in which there are sharp theoretical differences between MMT and orthodoxy. We will look at the case of minimum wage laws as an example of the difficulty of determining the macroeconomic effects of policy. Finally, we address the nature of scientific progress in the social sciences.

In the Appendix, we provide a brief outline of the **Buckaroos model**, which has been implemented at the University of Missouri at Kansas City (UMKC) in the United States. UMKC students are required to undertake a certain number of hours of voluntary labour for community service providers prior to graduation. The Buckaroos model is a means of operationalising the administration of this scheme and provides insights about the operation of a modern monetary economy.

2.2 Thinking in a Macroeconomic Way

Macroeconomics is a controversy-ridden area of study. In part, this is because the topic of study is seen as being of great significance to our nation and our daily lives, even though the details that are discussed are mostly difficult for us to understand.

The popular press and media in general are flooded with macroeconomics. The nightly news bulletins invariably feature commentators speaking about macroeconomic issues, such as the real GDP growth rate, the inflation rate or the unemployment rate. The population has been more exposed to macroeconomic terminology over the last two decades or so, and the advent of social media has given a voice to anyone who wants to be a macroeconomic commentator.

The so called 'blogosphere' is replete with self-appointed macroeconomic experts who wax lyrical about all and sundry, often relying on intuitively logical arguments to make their cases. The problems are that common sense is a dangerous guide to reality, and not all opinion should be given equal privilege in public discourse. Our propensity to generalise from personal experience, as if that experience constitutes universal truths, dominates the public debate; and the area of macroeconomics is a major arena for this sort of problematic reasoning.

A typical statement that is made in the public arena is that 'the government might run out of money if it doesn't curb spending'. Conservative politicians who seek to limit the spending ambit of government often attempt to give this statement authority by appealing to our intuition and experience.

They draw an analogy between the household and the sovereign government to assert that the same microeconomic constraints that are imposed on individuals or households apply equally, without qualification, to the government. We are told that governments, like households, have to live within their means. This analogy resonates strongly with voters because it equates the amorphous, obscure workings of government finance to a more easily relatable subject and scale – that of our daily household finances.

As we noted in Chapter 1, we know that we cannot run up our household debt forever and that we have to moderate our spending when we reach the borrowing limits on our credit cards. We can borrow to enhance current spending, but eventually we have to sacrifice spending to pay the debts back.

Neoliberals draw an analogy between households and governments because they know we will judge government deficits as being reckless, more so if fiscal deficits rise. But the government is not a big household. It can consistently spend more than its revenue if it creates the currency. Whereas households have to save (spend less than they earn) or borrow to spend more in the future, governments can purchase whatever they like whenever there are goods and services for sale in the currency they issue. Governments such as Britain, the United States, Japan and Australia always have the capacity to spend in their own currencies, and can never run out of money.

MMT teaches that our experience in managing our own household budgets provides no guidance about the management of the government fiscal position, yet on a daily basis, we are told it does.

The government has to consider the real resources that are available to the economy and how best to deploy them. These are not financial considerations; there are no intrinsic 'financial' constraints that are relevant to a currency-issuing government. In addition, fiscal surpluses (taxation revenue greater than government spending) today do not provide greater capacity to governments to meet future spending needs, nor do fiscal deficits (taxation revenue less than government spending) erode that capacity.

A household that has 'too much debt' can save and reduce that debt. But putting aside the question of whether public debt is actually inherently problematic (see Chapter 22), if the government tries to 'save' (in itself another inapplicable conceptual transfer from the individual level) then public debt will probably rise.

Prior to the 1930s, there was no separate field of study called macroeconomics. The dominant neoclassical school of thought in economics saw macroeconomics as a simple aggregation of the reasoning used at the individual unit or atomistic level. In the 1930s, macroeconomics emerged as a separate discipline precisely because that way of thinking, blithely transposing microeconomic truisms to the macro scale, was realised to be riddled with errors of logic that led to spurious analytical reasoning and poor policy advice.

Microeconomics develops theories about individual behavioural units in the economy – at the level of the person, household, or firm. For example, it might seek to explain the employment decisions of a firm or the saving decisions of an individual income recipient. However, microeconomic theory ignores knock-on effects on others when examining these firm- or household-level decisions. That is clearly inappropriate if we look at the macroeconomy, where we must consider these wider impacts.

We have learned that macroeconomics studies the aggregate outcomes of the behaviour of all firms and households. The question is how we go from the individual unit (microeconomic) level to the economy-wide (macroeconomic) level. This is a question that the so-called aggregation problem seeks to address.

To make statements about industry, markets or the economy as a whole, neoclassical economists sought to aggregate their atomistic analysis. For reasons that will become clearer, simple aggregation proved to be flawed. The solution was to fudge the task and introduce the notion of a 'representative household' to be the demand side of a goods and services (product) market and the 'representative firm' to be the supply side of that market. Together the two sides bought and sold a 'composite good'. These aggregates were fictions and assumed away many of the interesting aspects of market interaction.

For example, if we simply sum all the individual demand relationships between price and spending intention we could form a representative household demand function.

But what if the spending intentions of each household or a segment of them were *inter*dependent rather than independent? What if a household changed their demand once they found out what the spending intentions of the next door neighbour were (the notion of 'keeping up with the Joneses')? What if the actions of one household impinge on the feasible choices of another? Then a simple summation of demands is inappropriate.

But these issues were not considered, and the representative firm and household used were just bigger versions of the atomistic units while the underlying principles applied to explain the behaviour of these representations were simply those that were used to explain behaviour at the isolated, individual level. Accordingly, changes in behaviour or circumstances that might benefit the individual or the firm are automatically claimed to be of benefit to the economy as a whole.

During the Great Depression, this erroneous logic guided policy in the early 1930s and the crisis deepened. At that time, British economist John Maynard Keynes and others sought to expose the logical error that the dominant orthodoxy had made in their approach to aggregation by highlighting several discrepancies with that model, including the paradox of thrift and the wage-cutting solution to unemployment. In that debate this orthodoxy on aggregation was shown to incorporate a compositional fallacy, which led to the development of macroeconomics as a separate discipline from microeconomics. Karl Marx had appreciated this fallacy in the mid-19th century but his contributions were largely ignored in the popular economic literature of the early 20th century.

Compositional fallacies are errors in logic that arise when we infer that something which is true at the individual level, is also true at the aggregate level. The fallacy of composition arises here when actions that are logical, correct and/or rational at the individual or micro level are found to have no logic (and may be wrong and/ or irrational) at the aggregate or macro level.

Prior to considering the paradox of thrift and fiscal austerity, let us consider two simpler examples, the first of which is non-economic and the second has economic relevance.

Consider a large crowd attending a sporting event. The stadium provides seating for all attendees. An individual spectator would get a better view of an incident occurring near the sideline by standing up. Would all members of the crowd get a better view if they all stood up? Clearly the answer to this question is no – they would then obscure each other's views.

Now let us consider an employee who loses their job on Thursday evening. On Friday morning they consult the vacancies advertised in the local newspaper and online and apply for suitable jobs. They also knock on the doors of all local employers to present a CV and inquire about a job. Within a week they have secured a new job, following their thorough job search. Would it be correct to argue that if all the unemployed searched as conscientiously for jobs, then the unemployment problem would be solved? The answer is no. To make the discussion simple, assume all the unemployed are qualified to fill the available job vacancies, but 100 workers are competing for 50 jobs. At best, 50 of these job seekers will remain unemployed, irrespective of how thoroughly they search for jobs. This topic is further discussed in *The Tale of 100 Dogs and 95 Bones* (Centre of Full Employment and Equity [CofFEE], *c.*2001) (see Box 14.1).

A contemporary example of the flawed reasoning that follows a fallacy of composition is the **paradox of thrift**, which is that while an individual can increase their saving if they are disciplined enough (a micro-level fact), the same reasoning does not apply at the macro level. By reducing their individual consumption spending a person can of course increase the proportion they save, and enjoy higher future consumption possibilities as a consequence. The loss of spending to the overall economy caused by this individual's adjustment would be small and so there would be no detrimental impacts on overall economic activity, which is driven by aggregate spending. But imagine if all individuals (all consumers) adopted the same goal at the same time and started to reduce their spending *en masse*? Surely this *would* impact sales and hence employment and income at the aggregate level. It is not so clear that after all adjustments are made, we would find that aggregate saving had risen. This is what Keynes called the paradox of thrift.

Why does the paradox of thrift arise? In other words, what is the source of this compositional fallacy?

The explanation lies in a basic rule of macroeconomics, which you will learn once you start thinking in a macroeconomic way: that **spending creates income and output**. This planned economic activity powers the generation of employment to produce the goods and services. Thus, adjustments in spending drive adjustments in total production (output) in the economy as firms react to higher (lower) sales by increasing (reducing) employment and output.

As a consequence of increased saving, total spending falls significantly, and as you will learn from Chapter 15, national income falls (as production levels react to the lower spending) and unemployment rises. The impact of lost consumption on aggregate demand (spending) would be such that the economy could plunge into a recession. Certainly, total saving will be less than individuals planned due to the fall in equilibrium national income. As we will see later, if poor sales due to an increased desire to save negatively impact on investment, aggregate saving would certainly fall.

By assuming that we could simply add up the microeconomic relations to get the representative firm or household, the mainstream economist consensus at the time assumed that the aggregate unit faced the same constraints as the individual sub-units.

During the Global Financial Crisis (GFC), the conservative reaction to increasing government deficits was to enact fiscal austerity measures by cutting government expenditure and/or increasing taxes, and to encourage nations to cut domestic costs in order to stimulate their export sectors via increased competitiveness.

If one nation does this in isolation while all other nations are maintaining strong economic growth, this strategy might have a chance of working. In a similar way, one individual saver might reasonably assume that changing their consumption choices would not cause a wider effect that could impact their income. But if all nations engage in austerity and cut their growth rates, then overall spending declines, and imports will fall across the board, as will exports. This is another example of a fallacy of composition.

It is the interdependence between countries via trade, as well as a fall in net government spending, that undermines the policy prescription in this case. It is also clear that not all countries can rely on export-led growth (to more than offset a decline in net government spending) since for every exporter there must be an importer.

MMT contains a coherent logic that will teach you to resist falling into intuitive traps and compositional fallacies. MMT teaches you to think in a macroeconomic way.

Keynes and others considered that fallacies of composition, such as the paradox of thrift, provided a *prima facie* case for considering the study of macroeconomics as a separate discipline. The above examples show that we must be very careful when drawing general conclusions on the basis of our own experience.

2.3 What Should a Macroeconomic Theory be Able to Explain?

Any macroeconomic theory should help us understand the real world and provide both explanations of historical events and reasonable forecasts as to what might happen as a consequence of set events, for example, changes in policy settings.

A theory doesn't stand or fall on its absolute predictive accuracy because it is recognised that forecasting errors are a typical outcome of trying to make predictions about the unknown future. However, systematic forecast errors (that is, continually failing to predict the direction of the economy) and catastrophic oversights (for example, the failure to predict the 2008 GFC) are indications that a macroeconomic theory is seriously deficient.

In this section, we present some stylised facts about the way in which modern industrialised economies have performed over the last several decades. These facts will be referred to throughout the textbook as a reality check when we compare different approaches to the important macroeconomic issues such as unemployment, inflation, interest rates and government deficits.

The facts provide a benchmark against which any macroeconomic theory can be assessed. If a macroeconomic theory generates predictions which are consistently at odds with what we subsequently observe, then we can conclude that it doesn't advance our understanding of the real world and should be discarded.

Real GDP growth

Real gross domestic product (GDP) is the measure of actual production of goods and services in the economy over the course of a particular period. We will learn how the national statistics offices measure it and how we interpret movements in real GDP in Chapter 4 when we study the National Income and Product Accounts (NIPA). For now, we consider economic growth to be measured by the percentage change in real GDP, and in that sense, it is one measure of the prosperity of a nation. We will learn that employment growth is also dependent on output growth and so a higher real GDP growth usually means higher employment and lower unemployment.

Table 2.1 shows the average annual real GDP growth rates by decade from 1960 for various countries. The sample of nations chosen includes three large industrialised European nations representative of the 'north' (Germany) and 'south' (Italy and Spain), all of which are members of the Eurozone; a European nation outside the Eurozone since its inception in 1999 (UK); a small open economy predominantly exporting primary commodities and with a relatively underdeveloped industrial base (Australia); and two large, non-European industrialised nations (Japan and the USA).

Several things are clear. First, real economic growth has been lower on average since 2010 than in the 1960s for each country. Second, the southern European nations (Italy and Spain) have clearly performed poorly in the most recent period. Third, the European nations within the Eurozone, including Germany, have performed relatively poorly since 2000. Fourth, Australia has generally performed better than the other nations in Table 2.1.

Table 2.1 Average annual real GDP growth by decades, per cent

	Germany	Italy	Spain	UK	Australia	Japan	USA
1960–69	4.5	5.7	8.6	3.1	5.0	10.2	4.7
1970–79	3.3	4.0	5.3	2.6	3.3	5.2	3.2
1980–89	2.0	2.6	3.0	2.7	3.4	4.4	3.1
1990–99	2.2	1.5	2.8	2.1	3.2	1.5	3.2
2000–09	0.8	0.5	2.7	1.9	3.2	0.6	1.8
2010–15	2.0	−0.5	−0.3	2.0	2.6	1.4	2.1

Source: Authors' own. Data from the various national statistical agencies.

Among the questions that our macroeconomic approach needs to be able to answer in a consistent fashion are: Why has real GDP growth on average slowed? Why is Australia's growth rate since 2000 superior to that of the other nations? Why have Italy and Spain endured negative growth in the period 2010–2015?

Unemployment

One of the stark facts about modern economies has been the way in which unemployment has evolved since the 1980s. While different nations have recorded different outcomes, the common thread is that unemployment rates have risen overall, and in most cases, have been sustained at higher levels for many years.

In Figure 2.1, the unemployment rates, or the percentage of willing workers who are unable to find work, are shown for the seven nations depicted in Table 2.1 from 1960 to 2015. Please note that the vertical scales are different.

Table 2.2 provides further information upon which to assess the historical behaviour of unemployment.

Figure 2.1 and Table 2.2 show that unemployment rose in all seven nations during the 1970s and persisted at these high levels well into the first decade of the new century. Unemployment rates in Japan have been significantly below those of the other nations shown (although they have also trended upward).

The data also show quite clear cyclical patterns; Australia being a clear, pronounced example of this. Unemployment was below two per cent for most of the early post-Second World War period and then rose sharply in the mid-1970s, continuing to rise as the economy went into a deep recession in the early 1980s.

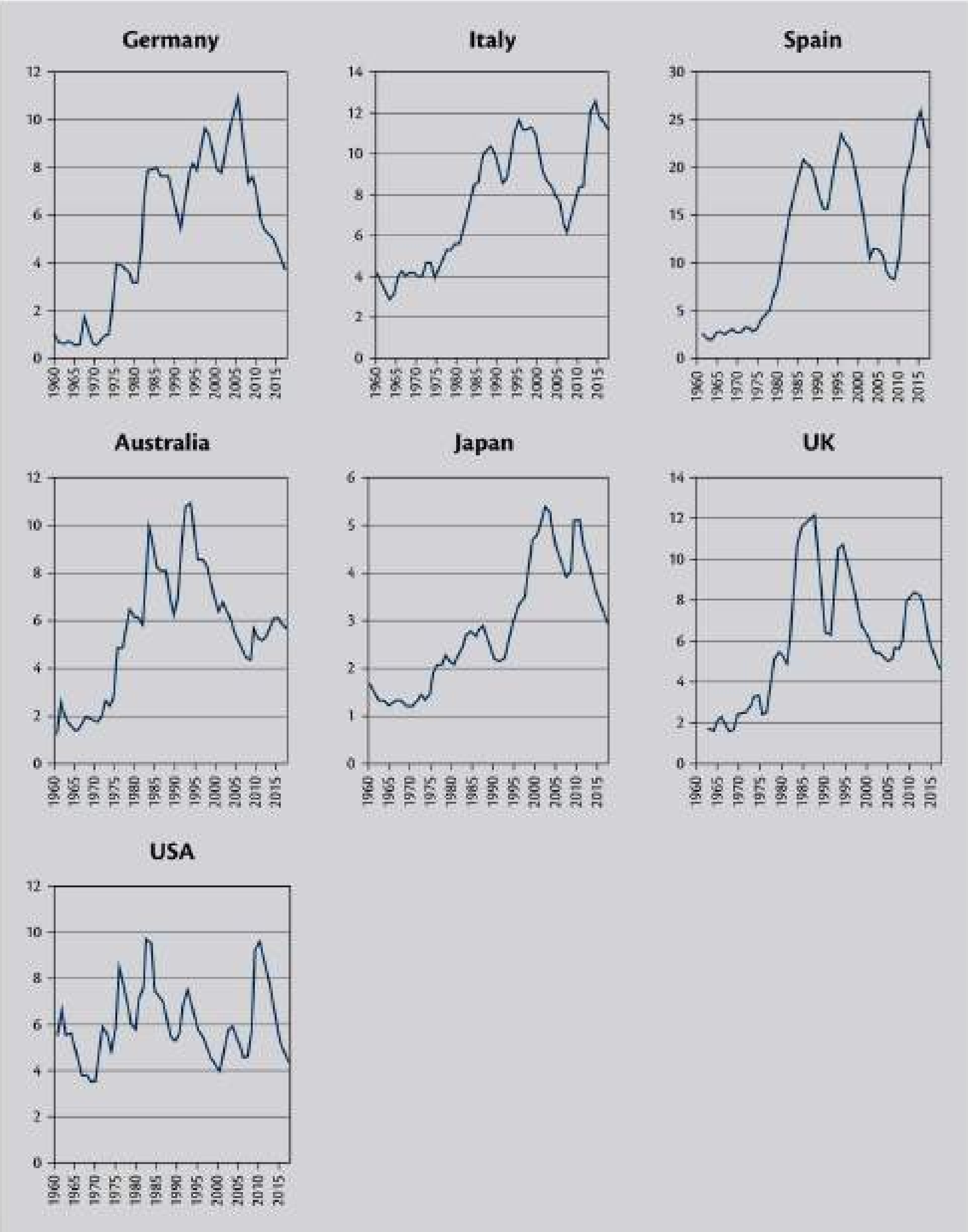
Economic growth in the second half of the 1980s brought Australia's unemployment down from its 1982 peak but never to the low levels of the 1950s, 1960s and early 1970s. The 1991 recession saw the unemployment rate jump up again very quickly and reach higher than the 1982 peak. The rate started to fall again as growth ensued after the recession was officially over, but it took many years to get back to pre-1991 levels.

Table 2.2 Average unemployment rates by decade, per cent

	Germany	Italy	Spain	UK	Australia	Japan	USA
1960–69	0.8	3.8	2.5	1.8	1.7	1.3	4.8
1970–79	2.4	4.7	4.4	3.6	3.9	1.7	6.2
1980–89	6.8	8.4	17.5	9.6	7.5	2.5	7.3
1990–99	7.8	10.4	19.5	8.0	8.8	3.0	5.8
2000–09	8.9	7.9	11.3	5.4	5.5	4.7	5.5
2010–17	5.1	10.9	21.9	6.6	5.6	3.9	6.8

Source: Authors' own. Data from the various national statistical agencies.

Figure 2.1 Comparative unemployment rates, per cent, 1960 to 2017



Source: Authors' own, using data from: Australian Bureau of Statistics; Federal Statistical Office, Germany; National Institute of Statistics, Italy; Ministry of Finance, Japan; National Statistics Institute, Spain; Office of National Statistics, Britain; Bureau of Economic Analysis, USA.

The US follows a similar pattern, although compared to Australia, unemployment rates were higher in the early post-war period but lower in the 1990s. The GFC largely bypassed Australia but led to high unemployment in the USA, which has fallen somewhat since.

Unemployment rates tend to behave in an asymmetric pattern; they rise very sharply and quickly when the economy goes into a downturn in activity but then only gradually fall over a long period once growth returns.

Any credible macroeconomic model needs to provide convincing explanations for these movements. How was unemployment kept at low levels during the 1950s and 1960s? Why did unemployment rates rise in the 1970s and persist at the higher levels for several decades? What determines the cyclical and asymmetric pattern of the unemployment rates? Is there a behavioural relationship between the GDP growth data shown in Table 2.1 and the unemployment data in Table 2.2?

In answer to the first two questions, MMT would refer to the key proposition in macroeconomics that total spending determines output and employment, and indirectly unemployment. MMT would conclude that the problem probably lies in insufficient spending, attributable to insufficient aggregate demand, a topic we will pursue in later chapters.

Real wages and productivity

In 1957, the renowned British economist Nicholas Kaldor wrote an article in *The Economic Journal* about the nature of long-term economic growth. He noted that there were six "remarkable historical constancies revealed by recent empirical investigations" (p. 591), which he later considered to constitute the 'stylised facts' regarding economic growth. He noted that these constancies were not necessarily immune to cyclical variation (as the economic cycle moves up and down), but were relatively constant over longer periods.

Among his stylised facts of economic growth was the observation that: "the share of wages and the share of profits in the national income has shown a remarkable constancy in 'developed' capitalist economies of the United States and the United Kingdom since the second half of the nineteenth century" (Kaldor 1957: 592–3). This observation was repeated by many economists for other nations in terms of the distribution of national income between labour (wages) and capital (profits).

We will learn in later chapters that for the share of wages and the share of profits in national income to remain constant over time, real wages must grow at the same rate as labour productivity. Real wages are the purchasing power equivalent of the wage a worker receives in money terms. Labour productivity is the output that is produced per unit of labour input.

Figure 2.2 picks up the story in early 1971 for the Australian and US economies. These examples are representative of trends observed over this time period in a number of advanced economies of the world. Up until the early 1980s, real wages continued to grow at the same rate as labour productivity (represented as 'GDP per hour worked' on our graph), which is consistent with Kaldor's observation.

In the early 1980s, a gap opened up between these two data series and has widened ever since. (Don't worry if you are having trouble interpreting the graph and its underlying data at this stage. As we work through the material in this textbook, we will develop the techniques necessary to assist you to interpret changes in data.)

In terms of shares of national income, the growing gap between real wages and labour productivity has meant that there has been an ongoing redistribution of real income away from workers (wages) towards capital (profits).

How do we explain this increasing discrepancy in national income shares? Why did the period of Kaldor's stylised constancy of national income shares end? What are the implications of such a substantial redistribution of national income away from real wages, which have until the last few decades been the primary driver of household consumption expenditure? What other factors now influence the growth in household consumption expenditure? This is a topic that macroeconomics must be able to explain, and this textbook will help you to address these questions.

Figure 2.2 Real wage and productivity indexes, Australia and USA, 1971 to 2015 (March 1982=100)

Source: Authors' own. Data from Australian Bureau of Statistics, National Accounts, US Bureau of Labor Statistics.

Private sector indebtedness

Figure 2.3 shows the rise in household debt as a share of disposable income between 2000 and 2015 for a range of OECD nations. Over the period shown, the ratio has risen markedly in most nations as financial market deregulation accelerated.

Is this large increase in the household debt ratio linked to the distributional shifts in national income implied by Figure 2.2? What other factors might explain this shift? What are the implications of the elevation in the household debt to disposable income ratio? Was the GFC linked to this movement? Again, this textbook will provide you with the understanding you need to comprehend these issues.

Central bank balance sheets

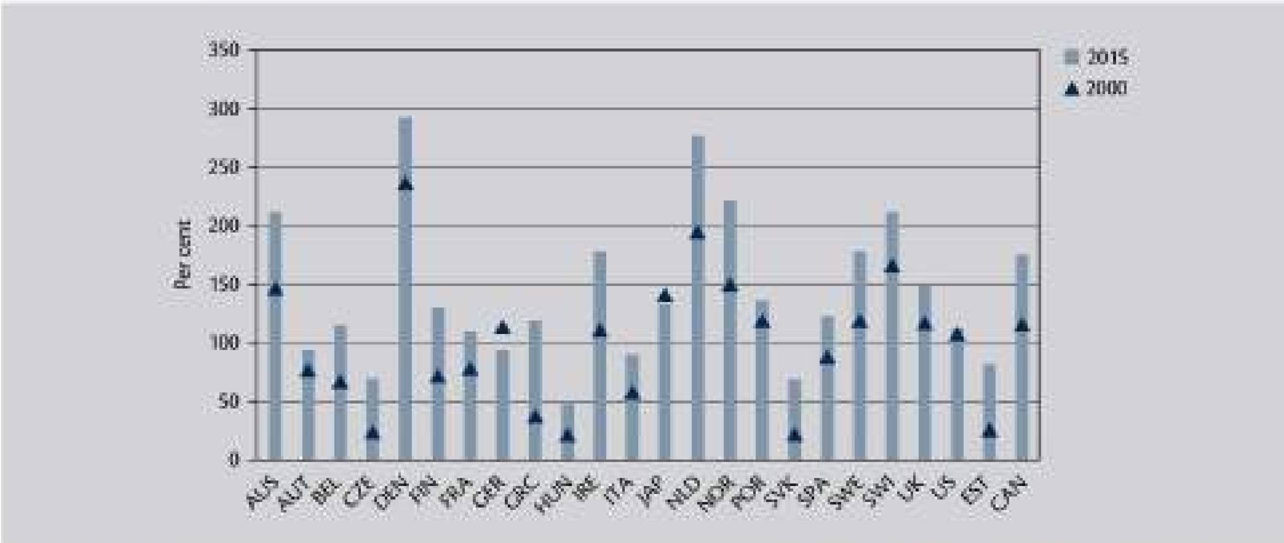
Figure 2.4 shows the so-called monetary base of the US economy administered by the Federal Reserve Bank. We will learn about the monetary base in later chapters but for now we can simply consider it to be the total reserves of the US banking system held at the central bank (the Federal Reserve Bank) plus currency (notes and coins) in circulation. The monetary base represents liabilities on the balance sheet of the US central bank. Up until 2008, the monetary base predominantly comprised currency on issue. The proportion in bank reserves increased from 2008 so that in December 2015 bank reserves were around 65 per cent of the total monetary base.

In January 2008, the US monetary base equalled $US830,632 million. It then accelerated upwards very quickly and by December 2015, stood at $US3,835,800 million – a huge increase by any standard.

The rise in bank reserves at the US central bank is not an isolated event and similar balance sheet shifts occurred in recent years in other nations (for example Japan and the UK). Many mainstream economists predicted that the substantial rise in central bank reserves would flood each economy with money and cause inflation. History tells us that over the same period inflation has been low.

How do we explain this massive shift in the balance sheet of the US Federal Reserve Bank? What are the implications of this shift? How does the monetary base relate to the money supply? Can the central bank carry liabilities of this size indefinitely?

Figure 2.3 Household debt to disposable income ratio, OECD nations, 2000 to 2015



Source: Authors' own. Data points from OECD iLibrary, *National Accounts at a Glance 2015*, Table 20.1. https://www.oecd-ilibrary.org/economics/national-accounts-at-a-glance-2015/household-debt_na_glance-2015-table31-en.

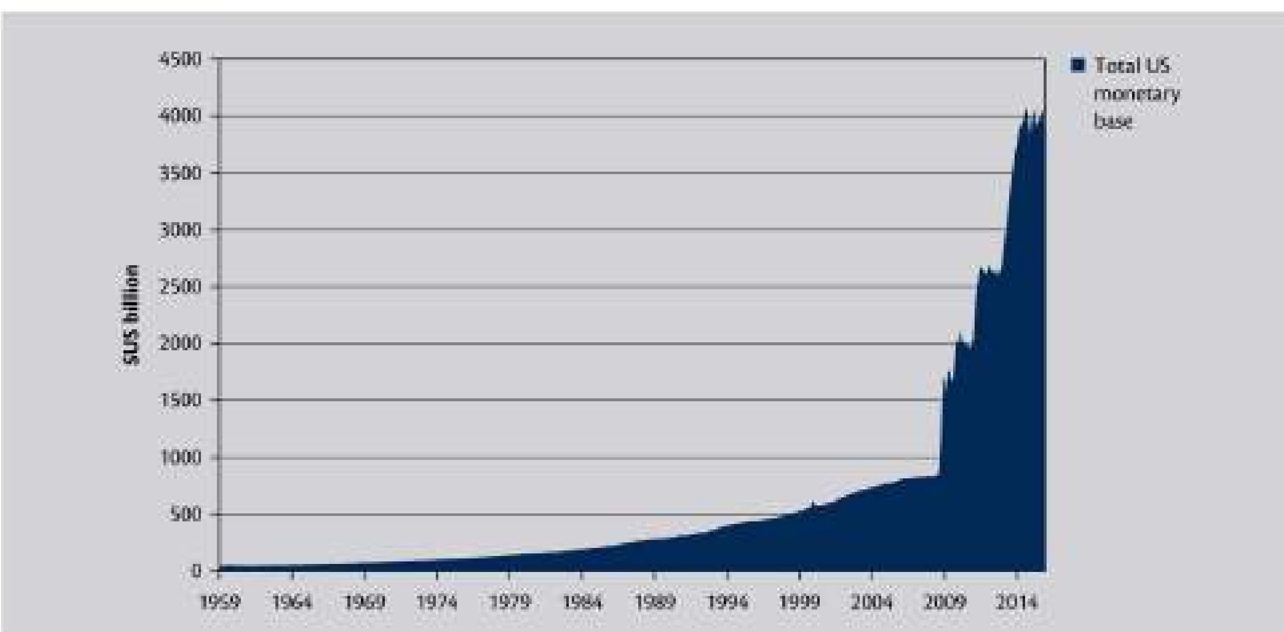
Notes: The abbreviations in the figure denote the following countries: AUS Australia; AUT Austria; BEL Belgium; CZE Czech Republic; DEN Denmark; FIN Finland; FRA France; GER Germany; GRC Greece; HUN Hungary; IRE Ireland; ITA Italy; JAP Japan; NLD Netherlands; NOR Norway; POR Portugal; SVK Slovakia; SPA Spain; SWE Sweden; SWI Switzerland; UK United Kingdom; US United States; EST Estonia; CAN Canada.

Japan's persistent fiscal deficits: the glaring counterfactual case

Consult almost any other macroeconomics textbook and you will find the following propositions stated, in some form or another, as inalienable fact:

1. Persistent fiscal deficits push up short-term interest rates because the alleged need to finance higher deficits increases the demand for scarce savings relative to its supply.

Figure 2.4 US Federal Reserve Bank monetary base, 1959 to 2015, $US billions



Source: Authors' own. Data from Federal Reserve Bank, US.

2. These higher interest rates undermine private investment spending (the so-called 'crowding out' hypothesis).
3. Persistent fiscal deficits lead to bond markets demanding increasing yields on government debt.
4. The rising public debt-to-GDP ratio associated with the persistent fiscal deficits will eventually lead bond markets to withdraw their lending to the government and the government will run out of money.
5. Persistent fiscal deficits lead to accelerating inflation and potentially hyperinflation, which is highly detrimental to the macroeconomy.

Japan was the second largest economy after its reconstruction following the Second World War led to spectacular growth in the 1960s. It is now the third largest economy behind the United States and China. Japan since 1990 provides a very interesting case study for macroeconomists because it has been marked by a number of macroeconomic outcomes that are at odds with orthodox thinking.

As we can see in Figure 2.5, Japan has run a persistent deficit since 1992. A massive build-up of private indebtedness associated with a real estate boom, accompanied the five years of fiscal surpluses from 1987 to 1991. The boom crashed spectacularly in 1991 and was followed by a period of lower growth and the need for higher deficits. The convention in Japan is that the national government matches its fiscal deficit with the issuance of bonds to the non-government sector, principally the private domestic sector.

Figure 2.6 shows the evolution of public debt levels as a percentage of GDP since 1980. Gross public debt is the total outstanding public debt issued by Japan's national (general) government sector. But the government also has investments which deliver returns, and when we subtract them from the gross public debt we get the net public debt.

Unsurprisingly, given the institutional practice of issuing debt to the private bond markets to match the fiscal deficits, the debt ratio has risen over time as a reflection of the ongoing deficits that the Japanese government has been running to support growth in the economy and maintain relatively low unemployment rates (see Figure 2.6).

If the neoclassical propositions summarised above correctly captured the way the real world operates, then we should have expected to see rising interest rates, increasing bond yields, and accelerating inflation in Japan, given the persistent fiscal deficits.

Figure 2.5 Government fiscal balance as a percentage of GDP, Japan, 1980–2015

Source: Authors' own. Data from IMF World Economic Outlook dataset (http://www.imf.org/weo).

Did the persistent fiscal deficits in Japan drive up interest rates and government bond yields? The answer is clearly no! Figure 2.7 shows the overnight interest rate in Japan, which is administered by the central bank, the Bank of Japan. This is the interest rate that banks use to borrow. It has stayed exceedingly low and has not responded adversely to the persistent fiscal deficits. Figure 2.8 shows that long-term (10 year) bond yields (interest rates) on government debt have also stayed very low and not responded adversely to the persistent fiscal deficits. If investors considered the government debt had become increasingly risky to purchase, they would have demanded increasing yields to compensate for that risk. There is no such suggestion – that bond market investors have become wary of Japanese government bonds – to be found here. Nor have they signalled any unwillingness to purchase the debt; demand for the bonds remains high and yields remain low.

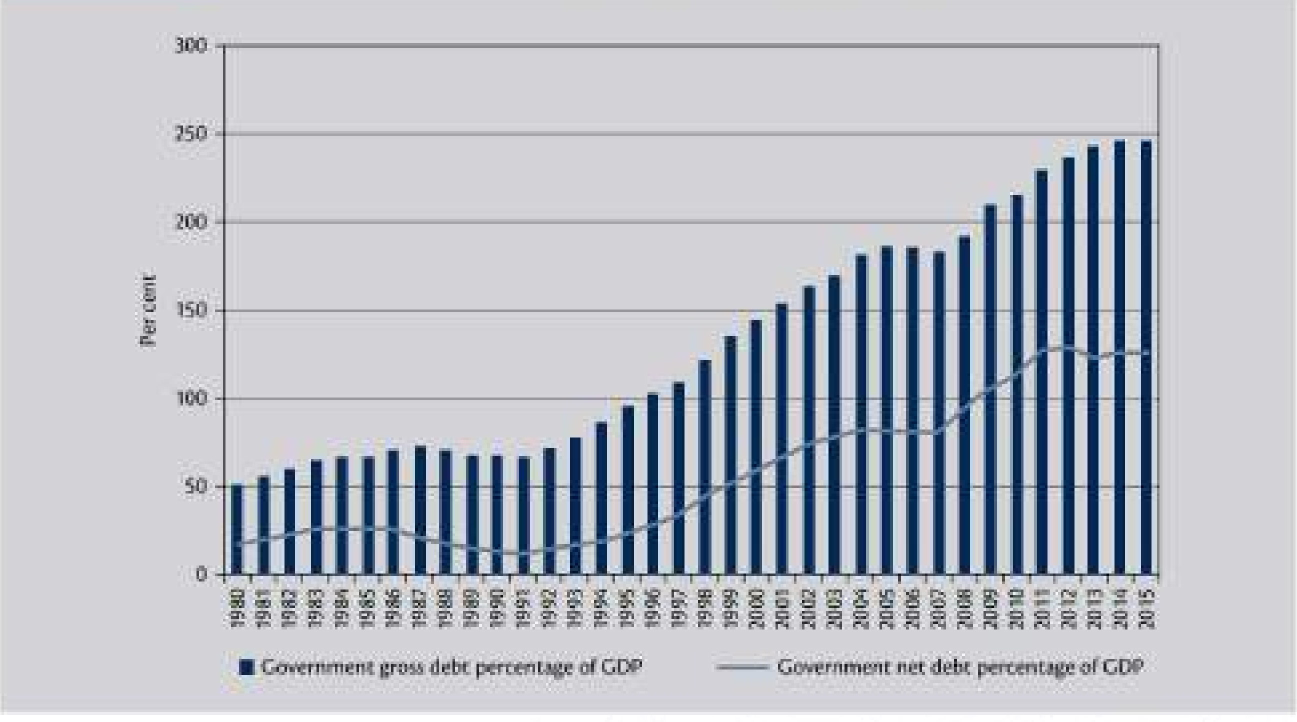
Figure 2.9 shows the inflation and deflation rates for Japan between 1980 and 2015. Inflation occurs when there is an ongoing increase in the general price level, whereas deflation describes the situation when the general price level is continuously falling (negative inflation).

You can see that since the property boom crashed and the Japanese government began to run persistent and at times, large, fiscal deficits, the inflation rate has been low and often negative. There is clearly no inflationary bias in the modern Japanese economy, as persistently predicted by the mainstream economic theories.

The above evidence shows that, despite persistent deficits and a rising public debt-to-GDP ratio, along with a downgrade of Japan's credit rating by international ratings agencies, including Fitch in April 2015, international bond markets have not 'punished' the Japanese government with high ten year interest rates on public debt nor has the central bank lost control of the overnight interest rate. Second, the persistent deficits have not led to high rates of domestic inflation.

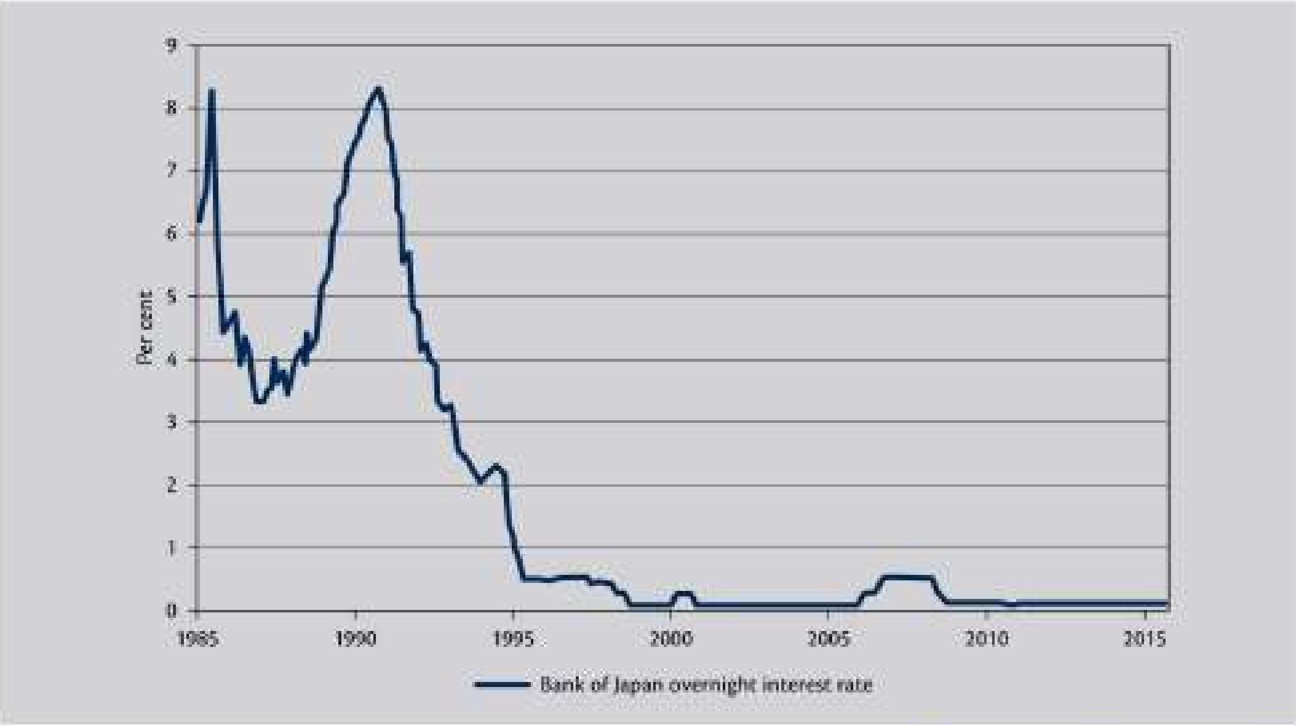
It is clear that the mainstream macroeconomic explanation of the relationships between fiscal deficits, interest rates, bond yields and inflation rates is unable to adequately capture the real world dynamics in Japan. Such a categorical failure to provide an explanation suggests that the mainstream theory is seriously deficient. A MMT

Figure 2.6 Gross and net public debt as a percentage of GDP, Japan, 1980 to 2015



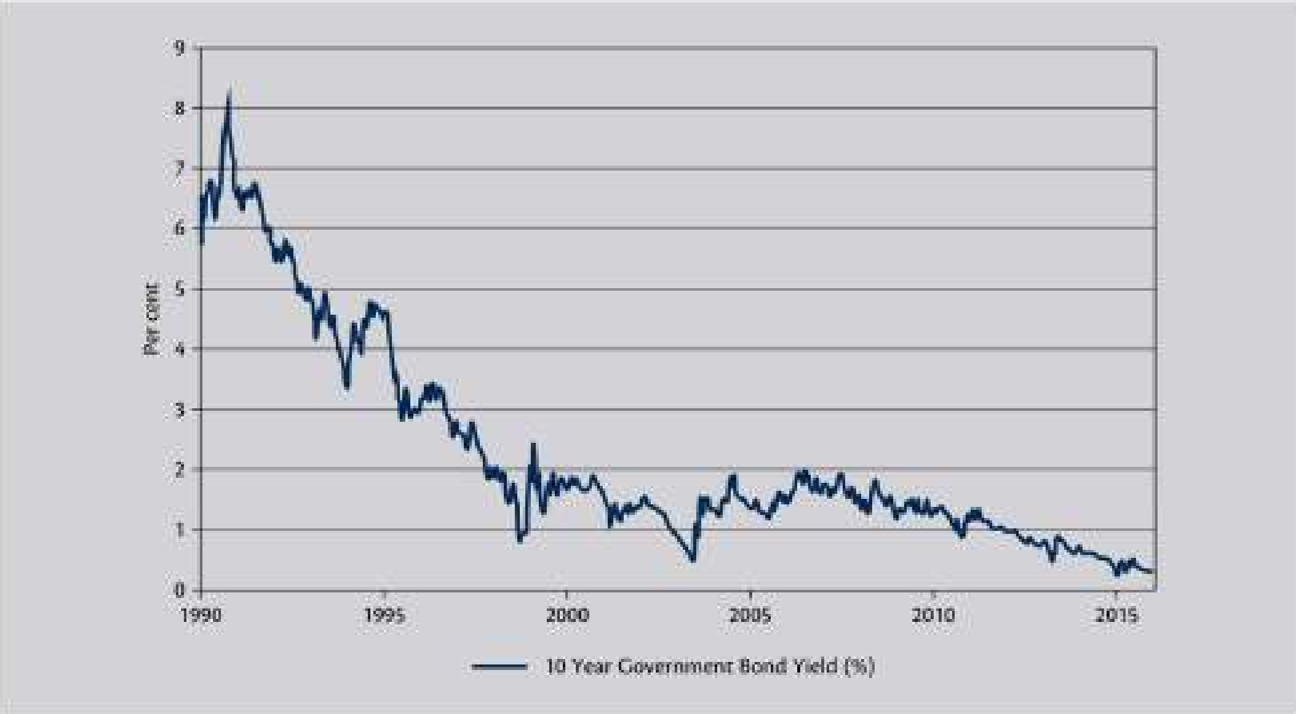
Source: Authors' own. Data from IMF World Economic Outlook dataset (http://www.imf.org/weo).

Figure 2.7 Japan overnight interest rate, per cent, July 1985 to December 2015



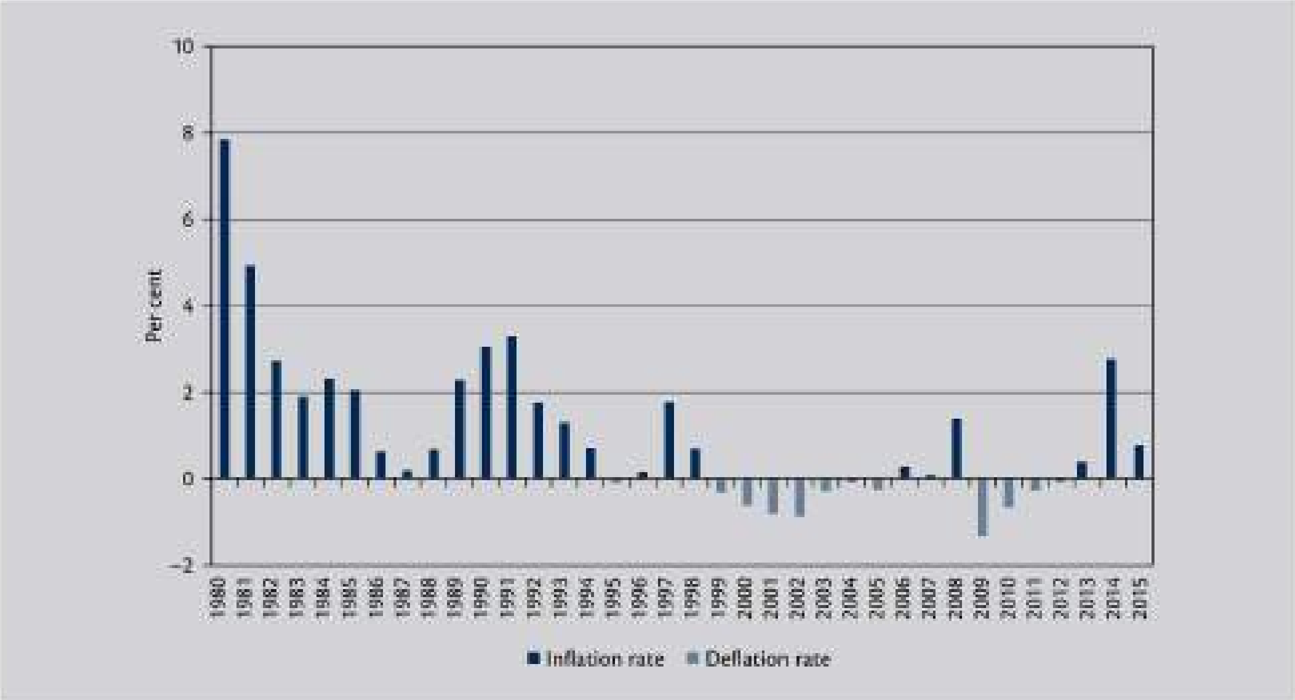
Source: Authors' own. Data from Bank of Japan (http://www.stat-search.boj.or.jp/index_en.html).

Figure 2.8 Japan government 10 year government bond yield, per cent, 1990 to 2015



Source: Authors' own. Data from Ministry of Finance, Japan (http://www.mof.go.jp/english/jgbs/reference/interest_rate/historical/jgbcme_all.csv).

Figure 2.9 Inflation and deflation in Japan, per cent, 1980 to 2015



Source: Authors' own. Data from IMF World Economic Outlook dataset (http://www.imf.org/weo).

explanation of these empirical outcomes will be provided in Chapters 20 and 21, when students will have developed a thorough understanding of the workings of a modern monetary economy with a sovereign currency, and the operation of fiscal policy.

2.4 Why is it so Difficult to Come to an Agreement on Policy? The Minimum Wage Debate

In the previous section, we explored some of the stylised facts and noted that these are things that economic theory ought to be able to explain. In this section, we turn to an example of a policy debate: the minimum wage.

Many countries have legislated minimum wages that are periodically raised. Each time policymakers examine the case for raising the minimum, economists are paraded to predict disastrous consequences for employment should wages be raised. Most of these economists dishonestly proclaim that economic theory reaches a decisive conclusion: minimum wages cause unemployment. In truth, economic theory gives (at least) two answers to the question of the effect of raising minimum wages on unemployment.

1. Raising wages increases business costs that beyond some point will increase the price of output. If we hold the income and purchasing power of consumers constant, it would seem that the higher prices must lead to fewer sales, and hence to lower employment. (There are other effects that could strengthen this impact, such as higher imports from abroad where labour is cheaper, and also substitution of machines for labour whose price has gone up.) Thus, neoliberals argue that raising the minimum wage must lead to higher unemployment.
2. Not so fast, says our two-armed economist. If wages rise, then it is not necessarily true that consumer income and hence purchasing power is constant. After all, most consumption is financed by wages, and the incomes

> of those employed at the lowest wages have increased. Those workers buy more goods and services. Firms which sell these goods and services might decide to hire more workers. Those workers buy more, too. If some employers decide that at the higher minimum wage they prefer to buy robotic machines to replace workers, that means more jobs making machines. We cannot say for sure that the net result of this complex chain reaction will be more jobs or fewer jobs.

So any two-armed economist will admit that economic *theory* cannot provide a decisive answer.

The frustrated student (and policymaker) asks 'But why can't we just look at real world evidence to settle the question?' Economists do of course try to do just that, and the tool of choice is econometrics. We can look at a number of cases where minimum wages have been raised. In the USA, for example, the 50 US states have their own minimum wage laws so it is possible to compare employment effects in one state when the minimum wage is raised while it is held constant in a neighbouring state with otherwise similar conditions. What the most careful studies in the USA find is that raising wages does not tend to reduce employment and raise unemployment; indeed it looks like the correlation goes the other way, with employment rising.

Does that settle the case? No. Even leaving aside clearly ideologically biased claims by opponents of minimum wage hikes, such empirical studies cannot be decisive. Even the most carefully designed tests cannot control for all possible factors that might affect employment. We cannot be sure that raising wages was what caused employment to rise. There could well be an uncontrolled factor that coincidentally increased employment (and indeed may have done so even if the wage hike by itself would have actually reduced the number employed).

Economists are well aware of this conundrum: **empirical correlation never proves causation**. Causation itself is a deeply complex topic. While we can put together theory and models and data to make a case, we probably will not be able to prove that 'X causes Y' when it comes to the most significant questions in economics.

John Maynard Keynes argued that the best one can do is to convince by the weight of one's argument. Certainly, one needs theory and probably evidence, maybe even a mathematical model, but even that will not convince an opponent unless the case is made through persuasive argument. Keynes was a master of argument, but even he did not always win. More recently Deirdre McCloskey made a similar claim in her book *The Rhetoric of Economics* (1985). Her point is that evidence alone is not decisive; 'rhetoric', the art of discourse, is also important.

If proof is difficult and theory provides ambiguous answers, can economics make progress? In the final section we address this question.

2.5 The Structure of Scientific Revolutions

In his influential book, *The Structure of Scientific Revolutions*, Thomas Kuhn (1970) advanced a thesis in which he distinguished between 'normal science' that works within a 'paradigm' and a 'scientific revolution' that breaks free of the paradigm – or even smashes it apart. For our purposes, we can think of the neoclassical approach as a 'paradigm' that works within the framework of utility maximisation and rationality, with the Keynesian (/Institutionalist/Marxist) approach as the paradigm-breaking scientific revolution.

Returning to the debate about the minimum wage, the neoliberal conclusion that raising wages causes unemployment is the correct answer if one views the question from within the neoclassical paradigm. In that paradigm, prices ration resources, and at higher prices there will be less demand. As wages rise, employers want fewer workers. It makes sense to argue that unemployment rises.

However, within the heterodox paradigm, what matters is aggregate effective demand (a topic to which we will turn later). Higher wages mean more income and more sales, hence firms want more workers. The net effect of a wage hike could be more employment.

Kuhn's breakthrough was the realisation that most of the time scientists (including economists) work within a paradigm, asking and attempting to answer questions in a manner that is consistent with the paradigm. He calls this 'normal science' and the research process mainly entails 'puzzle solving'. The 'normal scientist' comes across 'anomalies' that are hard to resolve within the paradigm in which they work, as we see in Box 2.1.

Kuhn's argument was that over time, as researchers pursue normal science working within their paradigm, they come up against more and more anomalies that cannot be explained. Another example would be the flat earth theory. Early scientists could come up with increasingly complicated explanations for the apparent anomalies. For example, as ships approach shore from a distant horizon, only the tops of the masts are first visible due to the earth's curvature. However, if light travels in a curved path that phenomenon could be explained within the flat earth paradigm. Yet, other tests would find that light apparently travels in a straight line, which is an anomaly.

According to Kuhn, as the anomalies build some researchers begin to think outside the paradigm. Well, perhaps the earth is *not* flat. Guests and servers may not be 'rational' in the narrow neoclassical sense. People begin to develop a new paradigm. Kuhn calls this a 'scientific revolution' and it has been likened to taking off distorting glasses and putting on prescription lenses that correct vision. The world never looks the same again because the new paradigm changes one's view completely. What were thought to be anomalies are easily explained within the new paradigm. It isn't a coincidence that the new paradigm is developed by younger researchers or by those outside the officialdom of the profession because it is easier for them to cast off the old ideas.

Within the new paradigm, normal science advances by puzzle solving, and eventually comes up against new anomalies. Eventually yet another scientific revolution will be needed. Note that no disparagement of 'normal' science is intended. Most of the advance of science comes through puzzle solving. Indeed, one cannot do research or even attempt to understand the world without a paradigm to start from. But puzzle solving, by itself, is not enough. Scientific revolutions are needed because paradigms are also constraining; they limit the conception of what is possible.

When he had finished the draft of his classic work, *The General Theory of Employment, Interest and Money* (1936), John Maynard Keynes wrote to a friend, George Bernard Shaw, proclaiming that his new book would revolutionise economic theory, if not at once, then at least eventually. That is quite a claim to make, of course, but Keynes was brilliant, and confident. The immediate reaction to his book seemed to validate his expectation. While not everyone was about to jump aboard it is not an exaggeration to say that many recognised

BOX 2.1 CHALLENGING NEOCLASSICAL CONVENTIONS

A commonly used example is the convention of tipping in a restaurant. If we assume that the diner and the server are both completely rational in the neoclassical sense [that is, selfish], then the tip typically should be negotiated before the meal to induce good service, except in the case where the diner is a local who often frequents the restaurant. The local diner can wait until after the meal to tip for good service. The server will provide good service in advance of the tip, expecting the diner will reward good service. If the local diner pays a low tip then poor service can be expected on the next visit.

The tourist or business visitor, however, might never expect to return to the restaurant. A tip before service could be negotiated depending on the level of service the diner wants. A contract is made and then if the server provides the service contracted the payment is made at dinner's end. The contract might include an external opinion and enforcement mechanism. In practice, we do not observe such contracts. Rather, the diner pays a tip at the end of dinner, based on assessment of services rendered. However, a rational one-time visitor would never pay a tip after service. Why bother? It is too late for the server to deliver poor service. And the diner never expects to return. Such behaviour is an anomaly for the neoclassical paradigm.

the revolutionary nature of his theory. By the 1960s, most macroeconomists considered themselves to be Keynesian.

And yet Keynesian theory soon fell out of favour. Mainstream macroeconomics began to shed Keynesian ideas from the early 1970s, and they were almost completely gone by the 1990s. It would be as if we returned to flat earth theory after once embracing round earth theory.

Note that part of the difference is that economics is a social science that studies human behaviour and proposes policy that directly affects human lives. It concerns topics that are contentious, and where policy benefits some but can hurt the interests of others. All the policies that came out of the Keynesian revolution were opposed by some groups, whether it was social welfare for the poor, social security for the aged, or jobs for the unemployed. Opponents inevitably regroup and attempt a counterrevolution.

Social sciences also experience reversals. Social theories from the past are thrust into the limelight again. Indeed, even in the 'hard' sciences, old ideas sometimes come back. In the USA, for example, the well-established theory of evolution is again under attack. Kuhn had warned that we should not see science as steadily progressing in a linear fashion from myth to truth. There is a tendency to write the textbooks in that manner, but reality is messy.

In any event, the authors of this textbook do view Keynes' *General Theory* as a scientific revolution, in Kuhn's sense, as were Karl Marx's theories presented in his 1867 book *Capital*. In both cases, orthodoxy mounted counterrevolutions to restore neoclassical thought.

By the 1870s three orthodox economists had published books to not only defend, but to strengthen, the arguments of neoclassical economics against Marx's economics. Jevons, Walras and Menger published their contributions between 1871 and 1873, in direct response to Marx (Henry, 2012). In other words, the neoclassical framework was developed as a rebuttal to the Marxian approach. Marx's revolutionary theory carried on, but in much of the West it was sidelined as the neoclassical theory became dominant.

In the case of the *General Theory*, the Keynesian revolution was gradually aborted as a few of Keynes' ideas were integrated into the neoclassical approach, forming the 'Neoclassical Synthesis' which was outlined in textbooks (with Paul Samuelson (1947) in the lead). All the revolutionary insights of Keynes (and Veblen and Marx) were dropped in order to make Keynes more or less consistent with neoclassical economics.

Unlike the case of Marx's *Capital*, which was openly disparaged, Keynes' book had been celebrated. A few of Keynes' ideas were incorporated into the 'Synthesis' and most macroeconomists became 'Keynesian' for some time, even though few fully understood the book.

Heterodox economists insist that this was a mistake, that neoclassical theory should have been dropped, and the revolutionary insights of heterodoxy (stretching all the way back to Marx) should have led to a new paradigm.

While our main purpose in this book is to develop the coherent heterodox alternative, we will present the neoclassical approach as we go along. Students must be familiar with both alternatives.

Conclusion

These examples demonstrate that macroeconomics is a highly contested discipline in terms of theory and policy prescription. When assessing the statements made by financial commentators and economists in the public debate, one must continually refer back to the stylised facts.

It is important that students gain familiarity with the language of macroeconomics and understand the key concepts and theories, which will be developed in the following chapters.

References

Centre of Full Employment and Equity (CofFEE) (c.2001) *The Tale of 100 Dogs and 95 Bones*. Available at: http://e1.newcastle.edu.au/coffee/education/education_view.cfm?ID=1, accessed 10 July 2018.

Henry, J.F. (2012) *The Making of Neoclassical Economics*, Routledge Revivals, Abingdon: Taylor & Francis.

Kaldor, N. (1957) "A Model of Economic Growth", *The Economic Journal*, 67(268), 591–624.

Keynes, J.M. (1936) *The General Theory of Employment, Interest and Money*, London: Macmillan.

Kuhn, T.S. (1970) *The Structure of Scientific Revolutions*, 2nd edn, Chicago: University of Chicago Press.

Marx, K. (1867) *Capital*, Volume I, London: Everyman's Library.

McCloskey, D. (1985) *The Rhetoric of Economics*, Madison, WI: University of Wisconsin Press.

Samuelson, P. (1947) *Foundations of Economic Analysis*, Cambridge, MA: Harvard University Press.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER 2 APPENDIX: THE BUCKAROOS MODEL

A modern monetary economy is characterised by a currency regime, whereby transactions between economic agents (for example, households, firms, financial institutions and government) can take place. This may involve, for example, the purchase of goods and services by households from firms; the purchase of assets by households and firms; the payment of taxes to the government or the receipt of transfers (e.g. unemployment benefit) from government.

The real world Buckaroos model demonstrates the roles of the currency, spending and taxes in a simplified economy.

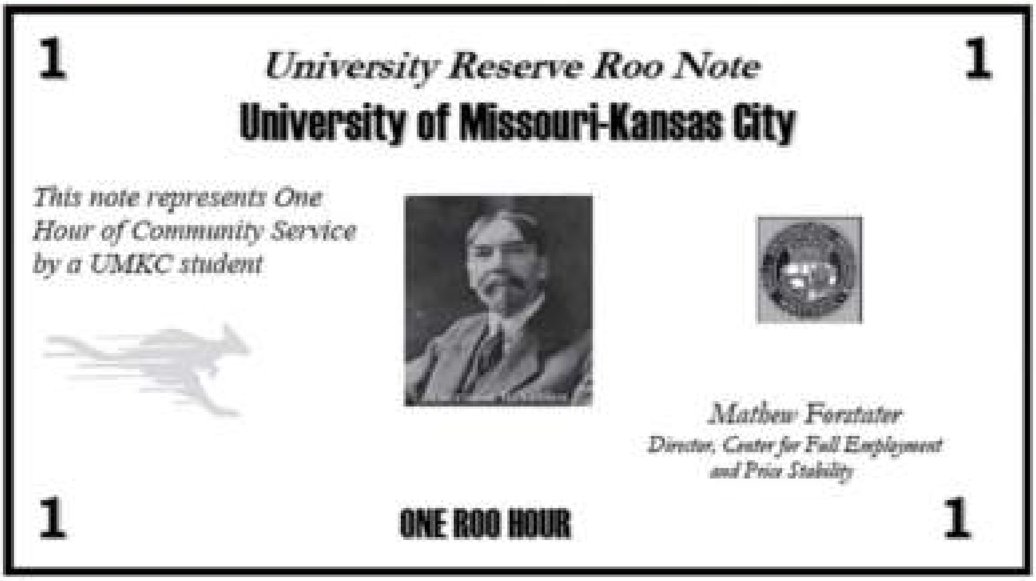
At the University of Missouri at Kansas City (UMKC), students are required to undertake a specified number of hours of Community Service (CS) during each year of their degree programme. Failure to complete the required hours of CS over the duration of the student's degree programme has negative implications for the final grade that the student receives. The Economics Department ran the pilot programme and designed a monetary system to administer the scheme, which we briefly outline below.

Each student is assumed to be subject to a community service tax of say 25 hours' work per semester, payable to the University Treasury. Assume there are University-approved CS providers (for example, child care, aged care, environmental services and so on) who submit bids for student hours to Treasury. Treasury awards paper notes (let's call these Bs as short for 'Buckaroos') to the CS providers (assuming health, safety and environmental standards are met). In this economy assume one hour of 'average community work' is equal to B1. Paper notes are printed, with the inscription 'this note represents one hour of community service by a UMKC student'.

For example, Treasury may agree that students can do a total of 100 hours of work this semester at 'the XYZ not-for-profit agency', which provides support for elderly people who are living alone. Treasury provides XYZ with B100, enabling 100 hours of student labour to be purchased.

CS providers then draw on their Bs to pay students for their hours of service. This can be considered 'spending' by the University Treasury, through the CS provider. If the student has undertaken 25 hours of CS in the semester, then they can then pay their B25 tax, when they return these Bs to the University Treasury. This transfer of Bs by each student to the Treasury extinguishes their tax liability for the semester.

Figure 2A.1 University Reserve Roo Note



The University Treasury burns the Bs received from students, or stockpiles them to be used for future treasury spending, whichever is more cost efficient. The number of Bs supplied to any CS provider is limited by its need for student labour, but also its ability to attract student workers.

Implications of the Buckaroos model

Treasury is the only source of Bs, which cannot be counterfeited. Treasury cannot collect B taxes until it has spent some Bs. Treasury can only be deemed to have spent when Bs are handed over to students for work done. Treasury cannot collect more Bs in payment of taxes than it has previously spent.

A possible Treasury outcome is a 'balanced budget', with tax 'revenues' equalling B spending. Thus Bs acquired by CS providers from Treasury are used to buy student labour and are then returned to Treasury as tax payments by the students. On the other hand, a surplus (deficit) arises in, say, semester one, if total treasury spending is less (more) than the total taxes collected over that period.

CHAPTER

3 A BRIEF OVERVIEW OF ECONOMIC HISTORY AND THE RISE OF CAPITALISM

Chapter Outline

Learning Objectives

- Recognise that the dominant mode of production evolves throughout history and can be overthrown.
- Recognise that while all societies face questions over production and distribution of the means of subsistence (including food, clothing, and shelter), they have created a variety of modes of production and distribution.
- Understand that the current dominant form around the globe is capitalism, but this system is relatively new, being only a few hundred years old. For most of human history, people have lived under other types of economic systems.
- Acknowledge that capitalism has evolved and may change fundamentally in the future, so there is nothing natural or everlasting about our current mode of production.

3.1 Introduction

In this chapter we will briefly examine what makes capitalism different from other economic systems. It is important to understand that humans have not always organised their economies around money. Throughout most of human history, the economy either operated entirely without money, or with money playing a relatively unimportant role in the provisioning process. However, with the rise of capitalism, it is not misleading to say that money came to play a dominant role. Hyman Minsky even argued that our modern capitalist system should be analysed as a *financial system*.

We should not, however, link the use of money only to the capitalist economy. Money has existed for at least 4,000 years, while capitalism's origins can be traced back approximately half a millennium. To be sure, money's origins are not really known and might never be known, but there is no doubt that money was used for thousands of years before capitalism rose to replace feudalism.

This chapter starts by noting that capitalism takes different forms. We briefly analyse different modes of production which have preceded capitalism, namely tribal society, slavery, and feudalism, and then examine the transition to capitalism. The repercussions of global capitalism are explored before we speculate about economic systems of the future.

3.2 An Introduction to Monetary Capitalism

Today the major nations have economic systems that conform to the general structure that is called capitalism. Sometimes these are inaccurately called **market systems**, a term that is both too general (markets predate capitalism by thousands of years) and too narrow (while markets are certainly important to capitalist systems, they are only a part of the economy). They are also called **mixed economies**, to indicate that the government sector as well as the private sector is important in the economic processes.

A somewhat more technical description used by some economists (including Marx and Keynes) is a **monetary production economy**; this captures the primary purpose of production for profit denominated in the money of account, which is a measure (denomination) of value used in keeping *accounts*. While that description does draw attention to the importance of money and the profit motive, it again seems to neglect the role played by government, which is not operated for monetary profit. For that reason, the simple term **capitalism** seems more appropriate for our purposes.

Marx considered the *differentia specifica* of capitalism to be its organisation of production, by which the owners of the means of production could secure surplus labour from workers who only had their labour power to sell in order to survive.

By surplus labour Marx meant labour that was performed in excess of the labour necessary to produce the means of livelihood for the worker or necessary labour. He thus distinguished the capacity to do work (labour power) from the physical act of working (labour).

It should not be thought that there is only one monolithic form of capitalism with uniform, carefully delineated institutions, rules of behaviour, and roles for government and other sectors. Capitalism takes a wide variety of forms, from a system comprised mostly of small-scale firms with a lot of the production farmed out to households employing simple tools, to a system utilising modern large-scale industrial production with literally thousands of highly skilled and unionised workers per factory. **Fordism** describes economic and social systems based on large-scale, standardised mass production and mass consumption.

Post-Fordism is now argued to be the dominant system of economic production and consumption, although again, definitions of the nature and scope of post-Fordism vary considerably. Commentators refer to features such as small batch production, specialised products and jobs, new information technologies and the growth of service employment.

Certainly, capitalist production has become increasingly *global*, with complex supply chains in which specialised plants located in a large number of countries each produce components that are finally assembled into the complete product.

Some capitalist firms might operate under the constraints of dog-eat-dog cut-throat competition, whereas others might be organised into large cartels that carefully control competition for mutual advantage. Capitalism can be mean, as depicted in the works of Charles Dickens, with most families eking out a miserable existence on low wages and long hours of work. Or it can be more generous, with a strong unionised workforce demanding good working conditions, adequate pay, and with a social safety net that takes care of aged persons, persons with disabilities, and children. And again, today capitalism is increasingly global, with relatively few unionised and comparatively well-paid workers in the rich nations while many of the component parts are produced in sweat shops with low pay and hazardous working conditions in developing nations.

Capitalist systems can perform well, offering rising living standards for most people, and they can collapse into great depressions as they did in the 1930s. They can grow fairly rapidly for long periods (Italy after 1960, or Japan until 1990) or they can stagnate with slow growth (Japan after 1990; most Western nations after the Global Financial Crisis of 2008 (GFC)).

Finally, capitalist systems can have 'big governments' that actively manage the economy to the benefit of the majority of the population, or they can have neoliberal governments that cater to the rich and powerful even as unemployment and poverty rates rise. Thus there are many capitalisms, and not one single kind of capitalism.

Finally, even if we recognise that there are many forms of capitalism, it is equally important to realise that capitalism is by no means the only kind of economic system. As we discuss in the following sections, humans have lived in other kinds of systems, and might choose to live in new systems in the future. We begin with a brief examination of other forms of economic organisation, before providing an outline of the development of capitalism.

3.3 Tribal Society

Historically, humans lived first (and for the longest period) in tribal societies. Both native Americans and native Australians, for example, lived in varieties of tribal forms of organisation at the time of being invaded by Europeans, as had all Europeans until the rise of Greece and Rome a few thousand years ago.

While recognising that there are many forms of tribal societies, we can generalise because they did share many characteristics, at least in the early stages of their development. First, tribes operated as egalitarian, communal, kinship-based social organisations. Members of a tribe were related by blood, with rules regarding marriage, initiation into adulthood (and adoption of new members who were not related by blood), and expulsion.

An egalitarian society is one in which members have equal rights and responsibilities, although there could be gender-specific and age-specific distinctions. It should also be noted that some tribal societies also practised slavery (generally, captured enemies were enslaved, killed, or adopted) so that egalitarianism did not apply to all those living within a tribe.

Typically, tribal society was matrilineal (one's heritage was traced through the mother's side) and matrilocal (upon marriage, the male joined the female's family) although there were also examples of patrilineal (lineage traced through father's side) and patrilocal (wife joined the husband's family) tribes.

Finally, a communal society is one in which production and distribution are undertaken by all members according to well-defined rules of participation. In other words, an individual (or individual family) would not be responsible for deciding what to produce, how to produce it, and when and how to consume the production. Instead, the tribe would decide what to produce, how to produce it, and how to distribute the fruits of the tribe's labour among its members. These decisions would follow custom, although adaptations would be made over time.

The great anthropologist Margaret Mead observed the complex rules adopted by a particular tribe regarding distribution of the meat from a hunt and of crops from farming. To modern Westerners it appears strange, but the hunter who killed game was not allowed to eat it, but rather had to turn it over to specific kin. The hunter's family would eat meat provided by other kin. This *reciprocal obligation* achieved the social purpose of binding members of the tribe together through mutual reliance.

There is some dispute over the possibility that tribal society made use of practices that might approximate market activity. It is doubtful that markets would have emerged within a tribe because production and distribution followed communal practice. An individual would not decide to produce something in order to sell it or to formally exchange it with another producer in the same tribe for some other item. Production and distribution decisions were made communally, so there would be no need to exchange essential items.

It is however well known that members of tribal society had elaborate ceremonies of gift exchange (something like Christmas gift giving), but the items exchanged usually had little practical value. The most likely purpose of these ceremonies was again to bring the members of the tribe together to enhance social relationships. Further, it was common to offer gifts to the family of the bride at marriage ceremonies, often called bride price. However, to view this as a market in which one buys a bride certainly seems to be erroneous.

Perhaps the tribal activity that comes closest to something that we might be willing to view as market exchange was the practice of gift giving between tribes. Some researchers have claimed that there are examples in which such exchanges involved the giving of useful items that the receiving tribe would not otherwise be able to obtain. For example, one tribe that lives in a rain forest might offer products that can only be produced from rain forest resources while another tribe that lives on grassland plain offers produce from its environment. In this case, the exchange of gifts is mutually beneficial in terms of improving living standards in each tribe, while also enhancing the social relations between the two tribes (reducing warfare).

Some might view this as akin to a moneyless market exchange, called barter. Even so, it is obvious that most provisioning is done within the tribe (not between tribes) and through communal production and distribution that does not involve markets. Thus, these exchanges were more likely to be ceremonial or symbolic than comprising a significant part of the participants' upkeep. There is little evidence for activity within tribal society that came close to markets with sales and purchases utilising IOUs denominated in a money of account.

3.4 Slavery

We noted that slavery existed in some tribal societies. There were also entire economies that were based on slavery, societies in which a large portion of the production of the essentials of life was done by the slaves. A relatively recent and well-known slave society existed in the southern states of the US until the Civil War of the 1860s. Students are probably also familiar with the slave-owning societies of Greece and Rome.

In slave societies, production decisions are mostly made by the owners of slaves, who also own the output of the slaves. Slaves can be bought and sold in markets, although there can be wide variations in the laws governing the treatment of slaves and their families. Typically, production by slaves is mostly used for consumption by the owner of the slaves and for subsistence of the slaves. However, slave production can also be used to provide goods and even services which are sold on the market.

Like tribal society and capitalist society, there are different forms of slave society. Some are much harsher in their treatment of slaves; some allow greater freedom for slaves or at least for their children. Some allowed slaves to gain freedom, with access to the human rights enjoyed by other members of society. The US version of slavery was particularly repulsive because it was combined with (and justified by) virulent racism that denied that blacks could even be considered human. By contrast, slavery in the ancient societies was not based on racism, meaning that freed slaves could gain rights of citizenship.

However, the most important point to recognise about slave society is that it is operated for the benefit of the slave owners (that are relatively few in number), and that the (typically) larger number of slaves recognise that their lives would improve through revolution and emancipation. Even in the most enlightened form of slave society, force is required to preserve slavery.

Further, because most of the benefits go to the owners of slaves, there is little incentive for slaves to increase productivity and expand output. Overseers are required to ensure at least a minimum work effort. Technological advance tends to be slow, both because slaves have little incentive or opportunity to innovate and because more complex means of production are typically easier to break and more costly to repair. Slaves could get back at their masters by throwing a 'monkey wrench' into the gears.

Slave societies are inherently weak and subject to revolt. (Many students will recall the name of the most famous slave to lead a revolt, Spartacus.) When faced with a military invasion, slave societies typically cannot arm slaves out of fear that the weapons would be turned against slave owners. Further, slaves are likely to use the opportunity of the invasion to initiate their own revolt. Thus, the state is unable to dedicate all its resources against the external threat, instead needing to maintain or even boost the forces retained to deter slave revolt. For these reasons, slave societies tend to be unstable.

3.5 Feudalism

Most students are at least passingly familiar with another important economic system, called feudalism. In Western Europe, the feudal period more or less coincides with the time period called the Middle Ages or Medieval period, but it is more accurate to use the term 'feudalism'. Knights and castles, lords and peasants, sword battles and jousting matches are all part of the Western lore which has been passed down for generations. Western and Eastern Europe as well as China and Japan each had a long experience with their own versions of feudal society. In Western Europe, feudal society emerged out of the fall of Rome in the 5th century BC, and lasted for a thousand years, although its institutions were beginning to break down by the 11th century, with a nascent capitalism beginning to replace feudalism from the 13th to the 17th century (depending on the region). As is always the case, one cannot put an exact date either on the beginning of an economic system or on its demise. And as always, there are different versions of the system we call feudalism.

For Western Europe, the most characteristic form reached its peak between the 7th and the 11th centuries. The two primary classes were the peasants and the feudal lords. The peasants had a right to agricultural land which was based on custom, with periodic redistribution of the land among the peasants to account for changes in family size, fertility of the land, and so on. Claims to the land typically went back as far as memories permitted, perhaps even to pre-Roman, tribal society times. Planting and harvesting were still typically done communally, as in tribal society, but each family would receive the output of the land to which they were assigned.

The right to farm the land should not be confused with modern conceptions of ownership of land, for a family was not free to **alienate** (sell) land. Indeed, a family could leave the region, and generations later an heir could return to claim a right to farm the land based on the family's ancient customary right, even after a market in land had developed. The institutions of transferable land ownership and private property were inconsistent with the feudal system, and once these developed, they would help to bring an end to feudalism.

The relatively small upper class consisted of the feudal lords. In some cases, they also had a customary right to land in the region. Because lords didn't work, peasants were expected to work the lord's allocated land, and to turn the produce over to the lord. The labour required to work the lord's allocated land was called rent.

Importantly, the surplus production (expropriated by the lord per his rights under the manorial system) was separated in time and space from the labour and production necessary to the worker to guarantee their own subsistence. This is a crucial step towards extrapolating labour into a commodity that could be bought and sold as under Marx's interpretation of capitalism.

In other cases, the lords did not have a right to land but, rather, exercised a right based on custom to a portion of the output of each peasant's land. This portion was also called **rent**, paid in the form of agricultural output. The knights, and other armed men who had sworn allegiance to the lord, enforced these customs through the threat of force in the service of the sheriff, who was charged by the lord to collect rent.

Over time, rents were gradually converted to money rents. Roman and Greek societies had money, and indeed had used coins. While these were scarce in the early days of feudalism, and were not necessary for the maintenance of feudal relations, as coins and other forms of money became more common lords would agree to accept them. Like the institution of private property in land, the growing use of money also helped to break down feudal society.

The payment of rent is often presented as an exchange, in which the peasants are buying protection from the feudal lord. In some respects, this was true. The (threat of the) feudal lord's armed forces did serve to dissuade and prevent other feudal lords from collecting rent from his peasants. However, the main protection that peasants received was from their own feudal lord and his knights, much as a storekeeper buys protection from a Mafia gang today. The payment of protection money ensures that the Mafiosi will prevent other gangs from using threats and violence to extort money. However, if the storekeeper doesn't pay, the Mafiosi themselves will assuredly break the shopkeeper's leg, threaten his family, or even set the store on fire. Similarly, if the peasant refused to pay rent, the knights would attack. If the peasant did pay, the knights would protect the peasant from attacks by other knights – the lesser of two evils. The peasants would have been happier without the need for protection from any of the feudal lords and knights, just as the storekeeper would be better off without rival Mafia gangs.

3.6 Revolts and the Transition to Capitalism

Again, just as in the case of slavery, it was not difficult for the peasants to conclude that they would be better off without the feudal lords. Hence, just as there were continual revolts by slaves in slave society, peasants revolted periodically against lords. The English Peasants Revolts took place in 1381, and were moderately successful at forcing concessions from the lords. The movie *Braveheart* depicts a 13th-century peasant revolt that was also nationalistic as peasants sought to drive out foreign lords and their armies.

By that time, many changes had already occurred in the nature of European feudalism. One important change was the increasingly common practice of paying rent in the form of money-denominated IOUs. Another change that was advanced by the 1381 revolts was the recognition of property rights in land, moving toward alienability of land. This, combined with the enclosure movement, helped to bring an end to the feudal system in Europe.

The enclosure movement is fairly well known to students of economics. Originally, a portion of the land in each region was preserved as a **commons**, containing forests, pastures, wetlands, and other land that was not farmed. The commons was important to peasant families as a source of wood for building and fires, for game, and for grazing cattle. However, feudal lords over time gradually exerted a claim to the commons, using the threat of force to keep the peasants out. The lords also claimed all the game and other resources of the commons.

The **enclosure** (that is, seizing) of the commons made an already difficult life truly unbearable for the peasants. They could no longer supplement their meals with game, they could not collect wood, and they could not graze their cattle. As a result, some would look for paid work to supplement their meagre output from farming. Some would even sell their land and abandon farming altogether. Wage receipts and receipts from the sale of land would often end up in the hands of the feudal lords, as peasants paid overdue rent.

As we can see, portions of the economy became increasingly monetised as feudal relations broke down. Working for wages became more common. Payment of rents in money form rather than in terms of labour or agricultural produce became more common. Land was bought and sold, displacing the customary rights to land.

At the same time, cities were becoming more important, acting as magnets for peasants who were leaving the land. In cities, one could perhaps find a position as an apprentice in a handicraft shop, learning skills producing furniture, silverware, or shoes. With luck and hard work, one might advance to a position as a master craftsman. Markets became increasingly important through specialisation. For example, the craftsman would produce shoes for market, and use the proceeds from sales to purchase food and other necessities, and perhaps even some luxuries. Peasants too could sell a portion of their output, paying money rent and perhaps purchasing some consumer items they had previously made themselves (or had done without). Markets and money became increasingly important.

In addition to the enclosure movement, other tactics were used to force peasants from the land. Some were chased off through violent attacks by the armies of lords and kings. Others were displaced by the seizure of the Catholic Church's lands. The Catholic Church was by far the largest feudal lord, exerting control over and collecting rent from vast areas. After a dispute with the Vatican, King Henry VIII confiscated the Church's lands in England. As a consequence, some peasants left their land voluntarily because they could no longer support their families, while others lost their land to creditors through excessive debt burdens.

In the 18th and 19th centuries, the Highland regions of Scotland were deliberately cleared of population, with great brutality, by the landowners who desired a greater income from their holdings. The vacated land could then be consolidated for pastures, particularly for sheep. The wool was shipped to the growing textile manufacturing industries. At the same time, the displaced peasants had to find alternative means of livelihood, so many were converted to wage labourers working in those same manufacturing industries.

There was a gradual transformation of the economy to something that looks a lot more familiar: employment, marketed output, cities, and even factories. Rather than an economy based on lord and peasant, we see workers and their employers, the capitalists. Even in the agricultural areas that were predominantly feudal, landowners increasingly employed wage labour (rather than peasants) to work the fields and tend the herds. A rising portion of output went to market.

Marketed output seeks profits, measured in terms of the money of account. While markets- and money-denominated sales, and money-denominated liabilities are thousands of years old, they had never dominated the economy previously. Most people's livelihoods over the previous centuries had not depended on producing marketable output. Most consumption had been satisfied by direct production by the consumer (or by their extended family).

With the breakdown of feudalism, all of that began to change rapidly.

3.7 Capitalism

The capitalist mode of production was altogether different from all previous economic systems. Following the development of capitalism, most producers (workers) had no right to the things they produced. They worked with tools and machinery owned by others – the capitalists. Indeed, most workers could not have produced much at all unless they worked for capitalists, because they had no other access to the necessary tools and machines. This is very different from feudal society, in which peasants had a customary right to the land that was necessary for agricultural production for subsistence.

The worker under capitalism has no *right* to the means of production, and hence no means of securing a livelihood *unless* they can convince an owner of the means of production (a capitalist) to employ them for wages. There is no guarantee that the worker will manage to obtain employment, and no guarantee that even if they do, the wage will be sufficient to produce an agreeable living standard.

While workers are sometimes called **wage slaves**, capitalism deviates in important ways from slave society. It is true that slaves in a slave society also had to work for somebody, generally did not own the tools they used, and did not own the output they produced. However, slaves were never unemployed. If an owner did not need a slave, the slave would be sold to someone who did. Further, a slave could not normally quit and search for a different owner. Workers are usually free to quit their jobs and to seek alternative employment; however, they can become unemployed, because they cannot force capitalists to hire them.

In contrast to the feudal system, in capitalism surplus production is not separated in time and space from the production required to maintain the subsistence for workers. Both are created by the day's work and the surplus accrued to the capitalists by dint of their ownership of the means of production rather than as a result of any manorial rights.

In mean capitalist systems, unemployed workers and their families would starve. In a slave society, by contrast, the rational owner of the slave would provide at least a subsistence level of necessities to protect their investment (their human property). Still, wage slavery (working for wages) is surely better than true slavery in which humans are reduced to the status of being the property of others.

All pre-capitalist societies experienced change through time: customs and beliefs evolved; technology, including know-how, changed and improved; animal power replaced human power and then was replaced by the power of machines. The types of things, and how they were consumed, changed; populations migrated; civilisations rose and then fell. Change has always been part of human life.

However, the pace of this change accelerated almost unimaginably under capitalism.

In previous forms of economic organisation, children could expect to live a life not noticeably different from the life led by their parents, their grandparents, or even their great-grandparents. Economic growth (in the sense of output of the production of the means of subsistence) generally occurred over time, but it was so slow that it was barely noticed. Capitalism changed all of that. While it is certainly not true to state that capitalism always and everywhere improved living standards, there was a nearly inexorable longer-term trend toward increased output that could not be overlooked.

In a very real sense, the whole discipline of economics was created with the rise of capitalism in order to make sense of this new, continually changing form of society. Often, the standard of living of one's children would be very different from that of their parents, and certainly the children would face a new array of products and jobs that could not have been imagined by their parents just 20 years earlier. The changes were obvious, and required the development of a field of social science that would explain the forces that drove them.

Capitalism represented a social revolution and indeed, the institutionalisation of mechanisms that ensure continuing change at a relatively rapid pace. It is very difficult for us to comprehend how rapid change is today when compared to the experience of all generations of humans prior to the rise of capitalism. There are people alive today in Australia and America who mostly travelled by horse-drawn cart and who lived in sod houses and who did without radio (and television!) when they were young. However, we should not be misled into thinking that living through rapid change is unique to the past half century; a similar pace of change was present even in the early days of capitalism.

The application of the term **industrial revolutions** (plural, because there were more than one) reflects the perceived pace of change, which was so fast and dramatic that it resembled revolution. Even those great critics of capitalism, Karl Marx and Frederic Engels, conceded that capitalism had **unleashed the forces of production** in a novel, unprecedented manner:

> *The bourgeoisie, during its rule of scarce one hundred years, has created more massive and more colossal productive forces than have all preceding generations together. Subjection of nature's forces to man, machinery, application of chemistry to industry and agriculture, steam navigation, railways, electric telegraphs, clearing of whole continents for cultivation, canalisation of rivers, whole populations conjured out of the ground – what earlier century had even a presentiment that such productive forces slumbered in the lap of social labour? (Marx and Engels, 1848)*

These words come from the *Communist Manifesto* (1848), which went on to proclaim:

> *The weapons with which the bourgeoisie felled feudalism to the ground are now turned against the bourgeoisie itself. But not only has the bourgeoisie forged the weapons that bring death to itself; it has also called into existence the men who are to wield those weapons – the modern working class – the proletarians. (Marx and Engels, 1848)*

Note how Engels and Marx celebrate both the rise of the monumental productive forces of **social labour** and declare that social labour (the proletariat) will prove to be capitalism's undoing.

Before we turn to the death of capitalism, let us look in some detail at the capitalist system in which most of the people of the world live today.

3.8 Monetary Capitalism

Capitalism is also very different from all previous economic systems because of its thoroughly monetary nature. We have hinted that money and markets long predate capitalism. However, what is important about capitalism is that the *purpose* of production is different. In all previous economic systems, the immediate purpose of production was to generate real goods and services to be consumed by the producers (members of the tribe, slaves, peasants) and by others (feudal lords and their knights, kings and queens, slave owners, ecclesiastical officials, and so on).

It is true that a portion of production was always dedicated to investment (in better tools, machines, and infrastructure) to ensure more efficient future production of goods and services. These deferred benefits were not available for immediate consumption by their producers. Also, depending on the working arrangements, some goods were produced for sale in markets even before capitalism. However, the majority of production never entered markets but rather was for the direct (even if not immediate) satisfaction of the tribe, the peasants and lord on the feudal manor, or the slaves and slave owner. And most of the goods and services consumed by the producers were not purchased in markets.

The capitalist form of production is undertaken to sell output and to make monetary profits. It is true that the produced goods and services *will* be consumed (or form part of the productive capacity to supply future consumption); however, this takes place only once they are sold.

Further, the production will not occur in the first place unless the capitalist, who owns the means of production, believes the goods and services can be sold at a *profitable* price. No matter how badly the population *needs* to consume or *wants* to work, the production will not occur unless it is deemed profitable.

Not only is money the goal of production it is also needed to allow production to proceed in the first place. Marx famously described the capitalist production sequence as $M \rightarrow C \rightarrow M'$. Production begins with **money** being used to purchase all of the inputs to the production process, including equipment, raw materials, and labour. This is represented by the initial *M*. The production process results in the production of **commodities** (*C*); these are goods and services that will be sold in the market. They include output that is sold directly to consumers as well as output sold to other firms to be used in other production processes. If all goes according to plan, the commodities will be sold at a sufficiently high price to reap profits. This requires that the total value of the **money received in sales** (*M'*) is greater than the total value of the **money used to engage in the production process** (*M*). That is, profits require that **M' > M**.

We can see that there are two main barriers to the production process that must be overcome: first, the capitalist must be able to obtain money finance (*M*) to begin production, and second, the capitalist must believe that sales of production will generate monetary profits (*M'* > *M*). Production can be prevented by either barrier.

Hence, in a *monetary production economy*, production begins with money on the expectation that the capitalist will end up with more money. In an important sense then, money can be blamed for unemployment, both of labour and of other resources. These resources will sit idle when capitalists either cannot obtain money to start the production process, or if they believe that production will not be sufficiently profitable (in terms of money).

3.9 Global Capitalism

As capitalism evolved, more of the production process was brought into the market. In the early days of capitalism, the family of the worker might still produce a large proportion of the goods it consumed. Milk, cream and butter would come from the family's cow; the garden produced vegetables; eggs came from the chickens. Much of their clothing and bedding would be made at home. Few services were purchased in the market.

However, there is a tendency for the capitalist form of production to continually expand into new areas in the pursuit of profitable opportunities. Today in modern developed capitalist countries, food mostly comes from global agribusiness, clothing is produced by large conglomerates employing cheap labour in Asia, and many of the services that households formerly performed for themselves are now bought in the market. For example, today in most US suburbs, even working-class families hire gardening firms to mow the lawns. This service would have been purchased only by the rich a few generations ago.

Not only does capitalism become more *intensive*, in the sense that it continually expands its reach into new markets within a nation, but it also becomes more *extensive* as it spreads over the globe and seeks to bring all potential consumers into the capitalist form of production. Beginning less than five centuries ago in northern Europe, the capitalist mode of production now dominates production almost everywhere in the world.

Further, as mentioned above, production increasingly takes the form of international supply chains. For example, the car parts that go to make up the typical final product are produced in factories all over the world. A major firm, such as Ford Motor Company will have many subsidiaries or even unaffiliated suppliers on contract producing parts in numerous countries, with the final assembly taking place in an identifiable Ford plant. The components of Apple's iPhones are largely produced by low-wage workers in Asia, although the high value software and marketing jobs are located in California. It is thus increasingly difficult to say exactly where 'production' takes place, which creates problems for regulators concerned with enforcing labour standards and for tax authorities. The rise of international supply chains adds force to the globalisation effect of capitalism as developing nations find it difficult to resist the demands of multinationals which might provide the jobs and development they want to attract.

However, we should not overstate the importance of capitalist production. Even in the most developed nations such as the US and Australia, much of the production that is absolutely essential for social survival takes place outside capitalist enterprises.

First, households still produce many of the goods and, especially, the services required to support the family: rearing children, cooking meals, routine maintenance of housing, gardening, financial services (balancing the family budget), entertainment (of and by each other), and so on. Even if much of this *could* be purchased in external markets, families perceive quality differences, and also enjoy working together. It is probable that healthy family life requires that a large portion of such activity be preserved from the reach of the market system.

Second, as we emphasise throughout this text, much of the production is better suited to public organisation and provision rather than to for-profit production. Since the 1970s, there has been a strong push by neoliberal politicians and think tanks to downsize government while either abandoning its responsibilities or contracting out services to private firms. This is justified by claims that private firms are more efficient and that the market produces the right incentives.

In some cases this is probably true, but in many others it opens the way for abuse, cronyism, and corruption. Further, since private firms are profit seeking, they rationally prefer to provide goods and services to those who are willing to pay and can afford to pay. For these reasons, there will always be room for production outside the market, by families, by government and by not-for-profit organisations to meet needs that are not fulfilled by for-profit production.

The push for globalisation has been very strong since the 1970s, as evidenced by various free trade deals. In the case of the US and its neighbours, an important recent development was NAFTA, which came into force in 1994. There was a similar period of globalisation at the end of the 19th century. In both cases, imports and exports became relatively more important, and huge international corporations took substantial control of international trade. In both the late 19th century and in recent decades, finance was also internationalised to a great extent.

That earlier period of globalisation and the associated development of international finance collapsed in the Great Depression of the 1930s. The US and other countries reacted by reforming finance, downsizing it, and exerting more control over it through regulation. International trade became somewhat less important, and trade barriers were restored, in the post-Second World War era. However over time, production, sales, and finance gradually became global, even more so than they had been in the early 20th century.

The GFC temporarily slowed the advance of global capitalism. However, the rescue of global financial institutions as well as of some ailing huge global non-financial corporations (such as General Motors) by the US Federal Government seems to have renewed its advance.

Globalisation became a major campaign issue in the US presidential election cycle of 2016, as both Bernie Sanders (from the left) and Donald Trump (from the right) attacked multinational corporations who are said to have moved jobs out of the USA to nations with cheap labour. With both Trump's win and 'Brexit', when the UK voted to leave the EU, many observers have predicted that the pace of globalisation might slow.

3.10 Economic Systems of the Future?

All economic systems evolve, but it is impossible to predict the direction of change. We can be sure that the economy will look different a hundred years from now, but we do not know how those differences will manifest. From the vantage point of the early 21st century, the form taken by capitalism in the major developed countries appears to be environmentally and socially unsustainable.

In later chapters, we will explore some of the social problems, particularly unemployment, inequality, and poverty, that result from the way that many modern economies function today. Undoubtedly, capitalism will continue to change, and informed policy can help to resolve these sorts of problems.

However, many critics of capitalism foresee a day when capitalism will be replaced by alternative economic (and socio-political) systems. We will briefly outline two such systems: socialism and communism. We will distinguish these from capitalist economic systems, and from each other.

Our definitions follow those usually used by the major advocates of such systems. However, it must be noted that the following is necessarily conjectural because we are describing possible future economic systems and we have no way of knowing how things will eventuate.

Finally, we realise that much controversy and confusion surround these terms. This is in part because several real world societies have variously claimed to be socialist or communist – or been accused by others of being socialist or communist – with varying degrees of accuracy. Here we set out clear definitions that are not meant to describe any of those real world economies.

A *socialist* economic system is one in which the means of production are collectively owned. In such a system, there are no functioning capitalists because private ownership of the means of production is prohibited. To be sure, there is still private ownership of clothing, of automobiles, of housing, and perhaps even of small family farms. However, a significant share of the means of production is not privately owned.

Without private ownership of the means of production, there is no significant private employment of other humans. Employment of family members or perhaps others within the household or on a family farm might be permitted. Most workers would be employed in organisations with collective ownership of the means of production.

Alternative arrangements would be possible. At one extreme, the workers of the collective would share communal ownership of the factory and all of its associated tools, buildings, financial assets, and so on. The collective would make all production decisions: what to produce, how to produce it, and how to price it.

At the other extreme, all means of production would be owned and managed by the workers of the nation as a whole. Decisions concerning what to produce, how to produce it, and how to price it would be made at the level of the nation as a whole by representatives of the workers. In this case, the purpose of production would be to achieve national goals. Unlike capitalism, production would not be undertaken with a view to obtaining monetary profits.

As a simple example of the difference between the systems, in a socialist economy, necessities (food, clothing, shelter, medical care, education) would be assigned sufficiently low prices that all members of society could afford them. Prices of luxury goods or of harmful products and practices (tobacco, gambling) would be set high enough to discourage their use.

By contrast, in a capitalist economy, prices are set to ensure that capitalist owners of firms achieve their desired profits.

A *communist* economic system shares some of the characteristics of the socialist system: there are no capitalists and no private ownership of the means of production. A further assumption is that all aspire to be workers and the practice of privately employing other humans is abhorrent to all members of society.

All production decisions are made democratically. Unlike socialism, there is no need for wages or prices because all production is freely and universally available to all. The forces of production are so great that all material needs and desires are easily satisfied. Hence, there is no reason to ration output.

Further, many of the social problems that spur conspicuous consumption and invidious distinction will have been removed from society. Hence, the sort of profligate consumption ('shop till you drop') that is common, or even tacitly encouraged, in developed capitalist economies will have disappeared. Rather than shopping malls and glossy advertisements that try to lure families to consume more than they need or really want, there will be communal warehouses at which families can obtain whatever they need.

In addition, the threat of deprivation will not be needed in order to induce people to work. All will want to contribute to a society that provides for their every need, and hence all will voluntarily participate in the social production process to the best of their ability.

There is a very simple way to distinguish between socialist and communist societies. Socialism provided the transition from capitalism to communism and in that system "the individual producer receives back from society – after the deductions have been made – exactly what he gives to it" (Marx, 1875: 10). In other words, the distribution of social output is largely determined by the contribution to the production process. This means that inequality of the distribution of output will continue under socialism: those who produce more will receive more. While the distribution will be less unequal than the distribution under capitalism, some inequality will remain. Of course, just as in other economic systems, there will be some who cannot produce very much. For example, people with disabilities, or people too young or too old to work, or parents with young children, might not be able to contribute very much to the production process. Thus, there will be some deviation from the socialist motto to ensure that all receive necessities.

The motto of communism is: "from each according to his ability, to each according to his needs!" (Marx, 1875: 11). In this case, there is no attempt and no need to ration output on the basis of the contribution to production. This is because the communist economic system can easily satisfy all reasonable needs, and the members of such a system will not have unreasonable desires. Each will take only what they need. Compulsion is not needed because each will contribute as much as they can.

If we compare either of these systems to capitalism, it is obvious that there are big differences. In a capitalist system, one's income includes earnings that are due to one's ownership of the means of production, which allows one to employ others and receive income generated by the production of others. The capitalist owners receive profit income not because they work, but rather because they own the factories and other establishments in which production takes place.

The capitalist system concentrates ownership of the means of production in the hands of a few, and then all others must work for the capitalist owners to generate profit income for them. While it is commonly claimed that capitalists also contribute to the production process by providing entrepreneurial skills, in practice these skills can be hired, there will be a hired management team, a hired research and development team, and so on. Even after paying all of these teams, there still must remain profit income or the capitalist owners will not allow their means of production to be used. What the capitalist owners actually provide is the means of production that they have effectively monopolised.

Socialism and communism eliminate capitalist income by eliminating private employment and private ownership of the means of production.

Conclusion

Will the economic systems of the future resemble the capitalism that is dominant today, the alternative systems studied in Section 3.10, or something new? We cannot know. There have been some real world experiments to implement socialist economic systems (the Paris Commune in 1871, the USSR in 1917, China in 1949). So far, it appears that none of these have been able to build a viable alternative to capitalism. The Paris Commune was crushed, the USSR collapsed, and China appears to be moving toward capitalism, albeit in a form that is rather different from that in the Western developed nations.

However, it must be remembered that the transition to capitalism required many false starts and several hundred years before it replaced feudalism throughout Europe. We can be sure that our economic system will continue to evolve, so that capitalism will likely undergo many transformations in the coming decades.

It also must be remembered that tribal society endured for perhaps tens of thousands of years, while feudalism persisted in Europe for about one thousand years. By comparison, capitalism is still a young upstart. And in spite of its problems, it has had a good run. So far.

The most important point we want to make as we conclude our brief introduction to economic history is that there is nothing natural about any particular form that our economy takes. Economic systems are *constructed* – even if not entirely intentionally so. And that means they can be changed if they do not perform well.

References

Marx, K. (1875) "Critique of the Gotha Program", in *Marx/Engels Selected Works* (1970), Vol. 3, Moscow: Progress Publishers. Available at: https://www.marxists.org/archive/marx/works/download/Marx_Critque_of_the_Gotha_Programme.pdf, accessed 4 April 2017.

Marx, K. and Engels, F. (1848) *Communist Manifesto*, English translation by Samuel Moore, 1888. Available at: https://www.marxists.org/archive/marx/works/download/pdf/Manifesto.pdf, accessed 3 September 2018.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

4 THE SYSTEM OF NATIONAL INCOME AND PRODUCT ACCOUNTS

Chapter Outline

Learning Objectives

- Understand how GDP is measured and why different measures of GDP are equivalent.
- Recognise the deficiencies of GDP as a measure of welfare.
- Derive CPI indexes and their rate of change.
- Interpret measures of income inequality.

4.1 Measuring National Output

The System of National Income and Product Accounts (NIPA) is the framework assembled by national statisticians for **measuring** economic activity.

In this chapter we look at national income accounting, that is, how we measure total national spending and its components as well as national income and its components. The most important measure of economic production is **gross domestic product** or GDP. Let us first provide a formal definition:

> **GDP is the measure of all final goods and services evaluated at market prices which are produced per period of time, say a quarter or a year.**

Note that GDP is a flow measure. A month, quarter, and year are the most common periods over which the flow of production is measured.

Let us emphasise the most important parts of the definition:

- **Produced per period of time**: This includes only goods and services produced over the time period, and would exclude goods sold this period that had been produced previously. Hence GDP excludes sales of 'used' goods.
- **Final goods and services**: These are the goods and services sold to final users, whether they are consumers, firms, or government. Households buy final consumer goods and services; firms buy investment goods to increase capacity, and government buys goods and hires services. Intermediate goods and services are excluded. For example, an auto manufacturer buys tyres to put on new cars for sale. These are intermediate goods. If we were to count those tyres as part of GDP, and then count the value of automobiles produced, we would double count the value of tyres (since the value of the automobiles would include all the intermediate goods and services that go into producing the automobiles). For that reason, we count only the value of final goods and services.
- **Evaluated at market prices**: We calculate the value of final goods and services at market prices. This means that GDP is calculated at nominal values. We use another measure of GDP to take account of the impact of price changes, called **real GDP**. Note that unless specifically designated as 'real GDP', when we say GDP we mean nominal GDP calculated at current market prices. We will discuss real GDP below.

The statisticians who compile the NIPA must make many decisions about what to include and what to exclude. While the decisions are not arbitrary, it is important to recognise that they are conventions. In other words, there is nothing sacrosanct about them, and the conventions could be changed by international agreement.

For example, washing your own dishes at home is not included in GDP. However, if you hire your neighbour to wash your dishes, that should be counted in GDP as dishwashing services. Note that we said 'should' because if you pay your neighbour 'under the table' and neither of you report it, the transaction might not get captured in the official numbers. This makes some sense because in the first case there was no monetary exchange and no market price at which the service took place, while in the second there is the market price that you paid for the service. However, by excluding all the unpaid household services performed, including cleaning, repairs and upkeep, and child and elder care, the NIPA numbers exclude a huge proportion of the nation's production.

More importantly, it undercounts the contribution made to production by women because they perform a disproportionate amount of unpaid work. Many economists have called for reform of the accounting conventions to include more unpaid work in order to give greater recognition to the economic and social value of such so-called 'women's work'.

GDP also excludes the black market, grey market, and much of the production in the informal sector. This is largely to do with the difficulty of collecting the data. Black market transactions are illegal, even though the good or service *per se* may be legal. For example, the sale of cigarettes is only illegal if duty has not been paid. On the other hand, the drug and sex trades often involves illegal transactions in many countries in illegal goods and services.

In the grey market legal, non-counterfeit goods are sold outside normal distribution channels. For example, if a brand of cameras is very expensive in a particular country, an enterprising local trader may import them from a country where the price is low and sell them in competition with the local official supplier(s) of the camera. Many nations do attempt to estimate such activity and even include at least some of it in official measures of GDP.

Much of the informal activity is similar to household production discussed above. For example, in many developing nations, much of the food production does not reach formal markets. It is consumed by farmers and shared or sold in local markets without being subject to proper recording. Other activity is 'under the table' and goes unrecorded to escape taxes. While the size of the black market is sometimes estimated in countries, typically it is not included in their official measures of GDP. However in late 2014 the Office of National Statistics in Italy announced that the estimation of its GDP would in future include the best estimates it could make for illegal activities, notably "drug trafficking, prostitution and smuggling services (cigarettes and alcohol)" (*The Economist*, 2014).

Another problem is that GDP is not necessarily a good measure of production in terms of its contribution to economic well-being. For example, a widget factory might pollute the air and water supply while it is producing its output. The social, health and environmental costs are not deducted from the value of the output produced for the purposes of measuring GDP. However, if society had to hire workers and produce machinery in order to clean up the pollution coming from the producing factory, that would be counted toward GDP. Ironically, this production would then count twice toward GDP, once for the value of the widgets produced and secondly for the value of cleaning up the environmental mess.

Furthermore, if neighbours of the widget factory get sick from the pollution, then the healthcare spending required to treat them also gets counted in GDP. For that reason, GDP can be a poor measure of economic well-being, as polluting industries might make a negative contribution to our general living standards even though they increase GDP.

Another example is when a nation that produces great volumes of military equipment might record the same GDP as a nation that produces a lot of healthcare and educational outputs. Using the GDP measure as a proxy for national progress might be quite misleading in this case. It is highly likely that the latter nation has higher material living standards for its population.

Still another problem is inequality. It does not make any difference to the calculation of GDP whether almost all production goes to the top ten per cent of individuals or households, so that the bottom 90 per cent gets next to nothing, or vice versa. The GDP measure simply adds up national production without taking account of the distribution of the output. This can make GDP a bad measure for comparing living standards across countries.

It is common to divide a nation's GDP by its population in order to derive the per capita GDP. We can rank nations according to their per capita GDP, measured in a common currency. We can classify some as rich, some as middle income, and some as poor. However, per capita GDP simply provides a measure of the average and that can be highly misleading as a guide to the standard of living of the typical resident of a nation.

For example, the average could be $35,000 per capita in two very different nations. In Country A, the share of GDP of the top one per cent might be 90 per cent, leaving the remaining 99 per cent of the population to share only ten per cent of the nation's output, while in Country B the distribution could be nearly equal, with 99 per cent of the population earning within a few thousand dollars of the $35,000 average. Clearly, economic well-being would be more widely shared in Country B, with very few poor people but also few people enjoying a living standard much above the average. The **Gini Coefficient**, which we discuss later in this chapter, is often used to measure income inequality.

There are alternative measures of economic well-being that attempt to get around these problems. Some try to measure household production. Others take account of inequality, poverty, and access to education and healthcare. Some measures deduct social, health, and environmental costs. For example, our hypothetical widget factory might actually make a net negative contribution to economic well-being, so it would be beneficial to close the factory and thereby increase social welfare even while forgoing the consumption of widgets.

As a real world example, tobacco smoking increases GDP due to sales of tobacco, spending to capture smoke to make indoor air cleaner, and high levels of spending on healthcare for tobacco users and all those who suffer from the effects of second-hand smoke. Eliminating tobacco use would undoubtedly enhance well-being but might reduce GDP. For these reasons, when addressing economic, social, and environmental well-being, we need alternative measures to GDP.

Still, GDP is the most commonly used measure and it does have one big advantage: it focuses largely on the monetary value of output. As we have discussed, the profit motive drives capitalistic production. It can be characterised as M-C-M'; that is, it begins with money (*M*) to produce commodities for sale (C) for more money (*M'*). For that reason, GDP is an appropriate measure for the capitalist sphere of production because it focuses on production for sale in exchange for money.

Yet GDP is not perfect even for that narrow purpose, because, as we have already noted, it includes imputed monetary values for some production that is not actually sold. The most important example is the 'services' of owner-occupied housing. The idea is that the homeowner 'consumes' housing services over time. If the home is not owned, we can use instead the rent paid as a value of the housing services consumed by the renter. However many families live in homes that they purchased so there are no market transactions taking place.

Note that when a new home is purchased, it is counted as residential investment (and is included in the investment category, not the consumption category – see the next section). It would not make sense to count the **entire** market value of a home as consumption over the period. Further, most homeowners have purchased a 'used' home, so that purchase will not show up in either the investment category or the consumption category. For that reason, the imputed monetary value of the housing services over the period is counted as consumption, whether or not the home is new. Still, by including imputed values, our measure of GDP deviates from the ideal of capturing the total value of production that is sold at market prices over the period.

4.2 Components of GDP

The National Income and Product Accounts (NIPA) divide the nation's output into four main categories, and add a fifth to account for foreign production that is available to the nation's residents. These are consumption, investment, government expenditure, exports and imports. Each of these can be further subdivided.

Consumption (*C*)

Consumption includes domestic consumption of goods and services by **households**. Keep in mind from our definition of GDP that only final outputs that are produced within a given time period – that is, currently produced final goods and services – are included. Intermediate goods and services are excluded, as are sales of used goods.

Generally speaking, all current period spending on new goods and services by households is included as consumption. The only major exceptions are the purchase of a newly built house, which, as noted, is included as investment spending, and the inclusion of 'imputed' housing services of owner-occupied homes, which is counted as consumption.

What is most confusing for students is that household 'investment' in shares and bonds is not included in GDP at all. This is because shares and bonds are not currently produced goods and services. Indeed, purchase of financial assets of any type is treated by the NIPA system as saving, not as spending.

Investment (*I*)

Investment includes three main categories: capital investment by firms, inventory investment by firms, and real estate investment by households. Investment expenditure increases the productive capacity of the economy and expands what we define as **potential GDP**. Therefore it adds to current spending but also increases the capacity of the economy to absorb increases in future spending without inflation.

Capital investment includes spending on plant and equipment; factories and machines, for example. Increasingly, investment includes purchases of software and other non-physical but long-lasting inputs to production.

As discussed, we do not want to include intermediate goods in GDP, so purchases of inputs that are 'used up' in the production process are not included as investment. Here we are referring to inputs such as electricity, oil and other natural resources, and marketing services. Note that the precise division between an intermediate input and an investment is somewhat arbitrary, and so it will rely on accounting conventions which will be related to the input's useful life.

Again, purchases of financial assets are not included as investment. For example, if one firm takes over another, that is not an investment for the purpose of measuring GDP. Also note that if a household buys a car it is counted as consumption, but if a business buys a car it is counted as investment, even if the firm operates out of a home office of the same household!

Unsold goods are referred to as inventories. A rise in inventories is also treated as an investment, even if the firm did not plan to change the stock of inventories. A firm may produce more output than it can sell during the accounting period, increasing investment in inventories. If a firm sells more output than planned, its stock of inventories falls. This is treated as negative investment. Swings of inventory investment can be quite wide because it is difficult for firms to sell precisely the amount that they planned.

Finally, real estate investment includes new construction of residential and non-residential buildings. Sales of existing homes as well as existing commercial buildings are not included as investment. Sales of land also would not be counted as investment.

When in doubt whether the purchase of an asset would be counted as investment or simply a purchase of an asset, a useful rule of thumb is to consider whether labour was used during the period to produce the asset. If it was, then this is investment; if not, then it is simply an asset purchase, which is treated as a portfolio adjustment, but not an investment. Newly produced machines, factories, houses, and apartment buildings all require current labour services to produce them, and hence count as investment. Sales of stocks, bonds, existing houses, or existing factories do not use labour in the current period to produce them, so they are not defined as investment.

Government spending (*G*)

Government spending includes government purchases of final goods and services.

Note that it does not include government transfer payments, such as spending on welfare and social security. This is because if we were to include transfers we would double count, since most transfer payments will then be spent on consumption goods and services, and hence, included in *C* as described above. Government transfer payments are not purchases of currently produced goods and services, so are not part of GDP.

Government purchases can be further divided between 'consumption' and 'investment' (or capital) expenditures. The division between these two subcategories is somewhat arbitrary. Government consumption expenditures are for goods and services that are used relatively quickly (firefighting services, postal delivery, and air traffic control), while government investment purchases are for long-lasting improvements (fire trucks, roads, and airports). Typically, any spending whose impacts are exhausted within a 12-month period are considered to be consumption; otherwise they are classified as investment. Do not get confused by the use of the terms 'consumption' and 'investment' when applied to the division of government spending by type; these are under the *G* category and not under the *C* or *I* categories discussed above.

Exports (*X*) minus imports (*M*) or net exports (*NX*)

Exports are goods and services sold abroad; imports are goods and services produced abroad for domestic use. If imports are greater than exports, then net exports are negative; alternatively, if imports are less than exports, then net exports are positive. Again, these can be consumption-type goods or investment-type goods but if they are sold abroad or bought from abroad they are counted in the *NX* category and not in the *C* or *I* categories.

Exports add to domestic spending to stimulate production, whereas imports represent a drain on domestic spending.

4.3 Equivalence of Three Measures of GDP

GDP can be measured in three ways, namely the expenditure approach, the production approach, and the income approach. These approaches, subject to a statistical discrepancy, should give equal measures of GDP.

The expenditure approach is conceptually the simplest. It works on the principle that total expenditures denote the value of the product that has been bought, and given the inclusion of inventory investment in the definition of investment, it measures the value of total production. The production (or value added) approach is based on summing the gross outputs of every class of enterprise and then netting out intermediate consumption. The income approach works on the principle that the incomes of the productive factors (producers) must be equal to the value of their product, and thus measures GDP by summing all producers' incomes.

Expenditure approach

The expenditure approach estimates GDP by calculating the sum of final expenditures on goods and services measured in current market prices. As we discussed above, GDP (Y) is the sum of consumption (C), investment (I), government spending (G) and net exports (NX or $(X - M)$). This can be written as the following identity:

$$Y \equiv C + I + G + (X - M) \tag{4.1}$$

An *identity* is an algebraic equation which always holds because of the way the variables in the equation are defined. Here we are drawing on the fact that the total final expenditure which represents GDP can be broken down into a series of components, which have been defined.

Production approach

This approach measures gross value added. First it is necessary to calculate the gross value of domestic output over, say, a year. This will include the value of output contributed at all stages of production,[1] and will take account of intermediate consumption – the costs of (raw) materials, supplies and services which were used up in the production of gross output. We then subtract the intermediate consumption from the gross value of domestic output to obtain the gross value added. As before, note that if we do not subtract the intermediate consumption, then we fall into the error of double counting.

Consider a three-stage production process which culminates in the final sale of woollen coats to consumers. Initially sheep farmers incur costs of feed and the like in rearing the sheep, and pay wages to the shepherds and to the sheep shearers. They then sell the wool to a woollen mill, which processes it by the employment of labour and other producers. The woollen mills sell the processed wool to the manufacturer of the coats, which employs labour and other producers in the production of the woollen coats. For simplicity, we assume that the manufacturer then sells these final goods directly to consumers. At each stage of the production process the value added by the producers must be calculated, so for example, value added by the woollen mill is the value of sales of the processed wool minus the costs of buying the unprocessed wool and raw materials used to process the wool and the electricity costs incurred in the production process. Then we can write: value added in the production of woollen coats = gross value of output – value of intermediate consumption, which has been summed over all stages of production.

The sum of the value added across every class of enterprise is known as GDP at factor cost. GDP at factor cost plus indirect taxes less subsidies on products is GDP at producer price.

Income approach

The third way of measuring GDP is to calculate the sum of primary incomes distributed to resident producers of goods and services. This method adds together the producers' incomes that firms pay in exchange for their services, namely wages for labour, interest for capital, rent for land and profit for capitalists. This defines GDP at factor cost. It is then necessary to add indirect taxes minus subsidies to get a measure at market prices, and in turn depreciation (or capital consumption allowance) to obtain GDP.

Under the production approach, the value added at each stage of production is the additional income that is generated, so the equivalence of the production and income approaches to the measurement of GDP is clear.

4.4 GDP versus GNP

GDP is the total value of goods and services produced within a nation regardless of the ownership of the firm producing them; gross national product (GNP) is the total value of goods and services produced by residents of the nation regardless of the location of the production.

GDP includes earnings from production within the domestic economy that goes to foreigners.

GNP does not include earnings from production within the domestic economy that goes to foreigners, but includes foreign earnings of domestic firms and residents operating abroad. Thus the financial flows between the domestic and external sectors are not confined to net exports.

Until the early 1990s the USA tended to use GNP while many nations used GDP. However, since then the USA has conformed and adopted GDP, although it still reports GNP. For the USA, there is no major difference between GDP and GNP because earnings from production in the USA that go to foreigners are nearly balanced against foreign earnings of US residents. For many other nations however there is a large difference between GDP and GNP because, for example, their residents have large investments in factories operating abroad. Ireland is a standout example of a country where these two measures diverge because of the presence of many large foreign-owned firms that have been attracted to locate head offices in that nation as a result of its very low corporate tax regime.

4.5 Measuring Gross and Net National Income

Measuring gross national income

We initially examine **gross national income** (GNI) from the perspective of what can be done with income: an individual can consume it, pay taxes, or save it. As a simplification we ignore the difference between GNP and GDP, so we can write:

$$Y = C + S + T = GDP = C + I + G + NX \tag{4.2}$$

Here we use Y to represent income. S is gross saving and T is total taxes paid. We can think of S as a residual: it is disposable (after tax) income that is not spent on consumption.

We can easily manipulate the above identity to obtain a useful identity based on Keynesian saving:

$$S = I + (G - T) + NX \tag{4.3}$$

What is $G - T$? It is government deficit spending, the difference between government spending (G) and its total revenue from taxation (T).

We'll make more use of identity (4.3) later.

Measuring net national income

At the aggregate level, national income equals national output because as discussed previously, the production of output generates equivalent income. We will define net national income as NNI, and then will derive a number of subcategories of income.

It is more convenient to begin with GNP (so as to include foreign earnings of domestic residents). GNP equals gross national income, or GNI. To calculate NNI we need to subtract depreciation and taxes.

Over the course of a production period (month, quarter, or year) some of the production facilities (plants and equipment) 'wear out' or 'depreciate'. We subtract depreciation from our gross national product (GNP) to obtain net national product (NNP).

We then subtract indirect business taxes (sales and excise taxes) to obtain NNI. The reason for deducting depreciation and business taxes is to obtain a measure of national income that is actually available to purchase national output. We subtract the depreciation because producers must set aside a portion of gross income to replace the capital that is wearing out. We subtract indirect business taxes because these reduce the amount of income that can be paid out of production. To summarise:

> Begin with GNP
>
> Subtract depreciation = NNP
>
> Subtract indirect taxes (sales and excise taxes) = NNI

Next we want to obtain a measure of personal income (PI) flowing to households. We subtract corporate taxes, payroll taxes, and undistributed profits since the taxes go to government and undistributed profits are retained by producers, leaving us with the income to be paid out to households.

However we need to add transfer payments made by government to households as well as personal interest income received by households to obtain PI. To summarise these operations:

> Begin with net national income (NNI)
>
> Subtract corporate taxes and undistributed profits and payroll taxes
>
> Add transfer payments and personal interest income = PI

We need to get a measure of PI after taxes paid by individuals, so we subtract personal taxes from PI to obtain personal disposable income (PDI). This is the after-tax income available to individuals to spend and save.

Beginning from PDI we then subtract personal consumption, interest paid to business and transfer payments made to foreigners to give us personal saving (PS).

> We start from personal disposable income
>
> Subtract personal consumption
>
> Subtract interest paid to business
>
> Subtract personal transfer payments to foreigners = PS

Note that gross saving (S) (defined in Equation 4.3) is not the same as personal saving (PS), because it is based on total income (not PDI) and we have not deducted interest paid to business and transfers to foreigners.

4.6 GDP Growth and the Price Deflator

We have defined nominal GDP as a measure of the value of output at current market prices. We often want to measure economic growth, which is the growth of GDP over time. The problem is that prices as well as output change over time. If we find that GDP (nominal) today is 100 times greater than it was a hundred years ago, does that mean that we enjoy 100 times more physical output? Clearly not if prices have also risen. To take account of this, we often want to 'deflate' nominal GDP, that is correct our measure for the change in prices to get an idea of 'real' economic growth.

The idea is simple, but in practice this is a very difficult thing to do. Let us start with the conceptual problem.

Suppose we want to compare GDP in 2018 to GDP in 2002 to see how much 'real' output grew over the 16-year period. To find nominal GDP in each year we take the 'current' market price of that year and multiply by the quantity produced that year. For exposition purposes, we are simplifying here by taking the quantity and price of a single aggregate good we call GDP:

$$(4.3a) \quad GDP_{2002} = P_{2002} \times Q_{2002}$$

$$(4.3b) \quad GDP_{2018} = P_{2018} \times Q_{2018}$$

where GDP_t measures GDP at current prices in year t, based on production level (Q_t) and market price (P_t).

However we are interested in a comparison of levels of 'real' GDP over time in order to correct our measure for the change in prices. First we have to decide which year's prices to use as a 'base'. We always calculate 'real GDP' over time in terms of a base year. We could choose 2002 or 2018 or any other year as the base. Let us say we choose to use the prices of 1985 (this makes it clear that we do not have to use prices of 2002 or of 2018).

Then we do the following calculation:

$$(4.4a) \quad RGDP_{2002} = P_{1985} \times Q_{2002}$$

$$(4.4b) \quad RGDP_{2018} = P_{1985} \times Q_{2018}$$

where $RGDP_t$ denotes real GDP in year t based on 1985 prices.

So long as we have used the same base year to calculate real GDP for both 2002 and 2018, we can determine real GDP growth over the 16-year period, but the measure will reflect to some degree the choice of the base year prices when we consider many goods rather than a single good.

In practice, statisticians update the base year through time so that they will always use a fairly recent base year. Thus, you would be unlikely to use 1900 as the base year to calculate real GDP for 2018! The older the base year used for calculations, the greater the problems encountered in calculating real GDP. We will return to these problems shortly. Before we do, there are two other useful concepts related to calculation of real GDP.

First there is the GDP deflator, which is an indicator of price changes. It is defined in year t as follows:

$$(4.5) \quad GDPD_t = GDP_t/RGDP_t$$

where $GDPD_t$ denotes the GDP deflator for year t.

Changes in the magnitude of the GDP deflator over time give us a measure of price changes for output as a whole. Note that it is possible for prices in general to go down as well as up. However, over the past century, deflations have been relatively rare and short-lived.

Our goal has been to develop a method for adjusting GDP for price changes. In practice it is much more difficult than suggested by the earlier discussion. As noted, we were using a simplification to calculate nominal GDP as 'Price times Quantity' of a single 'aggregate GDP' good.

However, GDP is defined as the value of total output measured at current prices. Conceptually we have a set (vector) of prices (one for each good or service sold) and a set (vector) of quantities (an entry for every item sold), and then we sum each individual sale ($P^i \times Q^i$ for the ith item) to obtain GDP. That does not seem too difficult; we simply recognise that output is heterogeneous and so it can only be aggregated in nominal terms, not in 'quantity' terms.

In practice, major problems are created if we try to measure the value of real GDP in terms of another year's prices. Let us say we use 1985 as our base year, and apply 1985 prices to the goods and services sold in 2018. How do we put a 1985 price on an iPad sold in 2018? There were no iPads sold in 1985, and indeed nothing comparable existed.

To reverse the problem, how can we find a 1985 price for manual typewriters sold in 1900 to value real GDP that year (in terms of 1985 prices)? Clearly, the composition of output changes both in terms of what is sold and the quality of items sold (the typical personal computer sold today is very much faster than one sold in 1990 even though the nominal price has hardly changed). It should be obvious that the older the base year chosen, the more acute the problem. That is why in recent years statisticians have increasingly favoured the use of a chain weighted measure of GDP which involves a lag of only one year. In the next section we discuss this measure in more detail.

4.7 Measuring Chain Weighted Real GDP

The chain weighted real GDP can be defined as follows:

$$RGDP_t = \{(P_{t-1} + P_t)/2\} \times Q_t$$

This measure averages the prices over two consecutive years and, as we discuss below, this is particularly useful for measuring real GDP growth.

In practice, economists are more interested in real GDP growth than in levels of real GDP. This favours the chain weighted measure even more over the calculation of real GDP with a base year that is periodically changed. Every time the base year is changed, real GDP needs to be recalculated for every year. That, in turn, will change the calculations of real GDP growth rates over time. In an important sense economic history is 'rewritten' every time the base year is changed.

With the chain weighted approach, however, the calculation of real GDP growth is invariant to changes of the base year. Changing the base year will change the calculated levels of real GDP but not the growth rate for the historical series of real GDP that will instead use the chain weighted measure.

Changes in this measure are calculated using the weights of adjacent years. These annual changes are 'chained' (multiplied) together to form a time series that allows for the effects of changes in relative prices and in the composition of output over time. Thus, the US Bureau of Economic Analysis (BEA) is able to calculate an index that uses weights that are appropriate for each period. It thereby avoids the rewriting of economic history that results from updating the base period of a fixed weighted index as well as the substitution bias that is inherent in fixed weighted indexes (Landefeld and Parker, 1997: 59–60).

In other words, once the BEA has calculated real GDP growth for any set of years using the chain weighted approach, it will not need to do any recalculations because the base year prices used for that set of years will not change. This is still more difficult than it sounds, but we will not go into further details here.

4.8 Measuring CPI Inflation

The CPI index

In this section we look at the measurement of the prices of consumer goods (bought by households) and make brief reference to producer goods (bought by firms, including raw materials and intermediate goods to be used in production). These prices could go down, but the usual trend is for rising prices.

The index most commonly used to calculate inflation of consumer goods prices is the Consumer Price Index, or **CPI**. It is defined as follows:

An index based on the cost of a fixed basket of consumer goods and services.

In the construction of the CPI index, the statistician needs to decide what consumer goods and services to include, their respective quantities (weights) and how to calculate the corresponding prices. The chosen basket of goods and services is intended to be representative of the purchases made by a typical household, and is periodically updated. The statistician chooses a base year (much like the choice of the base year to be used in calculating real GDP). The CPI then represents the cost of a market basket of consumer goods and services.

The measure is usually expressed for a specific spatial area such as a capital city or a weighted average of all capital cities in a nation.

The items included in the Australian CPI published by the Australian Bureau of Statistics in March 2016 are shown in Table 4.1. Within each major group there are many items included.

Table 4.1 Items in Australian CPI, March 2016

Food and non-alcoholic beverages
Alcohol and tobacco
Clothing and footwear
Housing
Furnishings, household equipment and services
Health
Transport
Communication
Recreation and culture
Education
Insurance and financial services

Source: Australian Bureau of Statistics, Consumer Price Index, http://www.abs.gov.au/AUSSTATS/abs@.nsf/mf/6401.0

If the prices of all the items in the basket changed at the same rate from one period to the next, then the change in the cost of the basket would be easy to calculate, period by period. But in reality, the individual prices generally change at different rates, so that relative prices are also changing. The statistician thus needs a single summary measure to determine whether the basket overall is rising in cost or not. That is the role that the price index plays. It is a weighted average of the price movements in the given basket relative to some base period.

In compiling a summary measure such as the CPI, the statistician must choose whether to use base weighting or current weighting to compile the index.

A base weighted index examines the shifts in prices of the basket of goods and services using the base period quantities purchased. It is referred to as a *Laspeyres* index after the German economist who first compiled such measures. The base weighted index allows us to see how much a basket that consumers bought in the base period would cost in the current period.

A current weighted index uses the current quantity purchased of each good and service in the basket as the weight to compile the average measure. This is commonly called a *Paasche* index after the German statistician who developed this measure. The current weighted index allows us to see how much a basket that consumers buy in the current period would have cost in the base period.

These measures provide different ways of estimating the change in the cost of a basket of goods and services over time. However, statisticians tend to favour the use of the *Laspeyres* index to calculate the CPI because it requires less information. The only new data that are required are the current prices of the items in the basket. The quantities making up the basket and the corresponding base year prices are already known.

This allows for a more timely publication of the CPI, which is desirable since it is an important policy variable used by central bankers and treasuries in formulating monetary and fiscal policy, not to mention its use in labour and other contracts, and for indexing the values of transfers, such as pensions and other benefits.

To simplify our analysis, imagine a basket of goods and services comprises two items: bread and cheese. (We are glossing over the obvious question of "Don't these people wear clothes?")

Table 4.2 shows the hypothetical data we will be working with to illustrate the construction of the price index.

Table 4.2 Hypothetical data for basket of goods and services

	Price per unit	Quantity	Expenditure	Expenditure based on quantities in other year
	(1) $	(2) Units	(3) $	(4) $
Year 1				
Cheese	4	3	12	16
Bread	2	9	18	20
Total			**30**	**36**
Year 2				
Cheese	5	4	20	15
Bread	3	10	30	27
Totals			**50**	**42**

In Year 1, the price per unit of cheese is $4 and three units are consumed overall. So total expenditure on cheese in Year 1 is $12. The price of a loaf of bread is $2 and nine units are consumed in Year 1, making total expenditure on bread $18. Overall, the basket of goods costs $30 in Year 1 (Column 3).

In Year 2, cheese rises to $5 per unit and four units are consumed whereas bread rises to $3 per loaf and ten units are consumed. Overall, the basket of goods in Year 2 now costs $50 (Column 3).

Note that if we wanted to know what the quantities purchased in Year 1 would cost in Year 2, we multiply each quantity by its price in Year 2. Column (4) provides that answer ($42), which we calculated using the following data:

Cheese $5 × 3 = $15

Bread $3 × 9 = $27

Total = $42

Conversely, Column (3) shows the expenditure in Year 2 based on Year 2 prices and Year 2 purchases.

Similarly, if we wanted to know what the basket would cost in Year 1 based on Year 1 prices and Year 2 purchases we would look to Column (4).

What would be the price index values in this example? Our answer will depend on whether we use base weighting or current weighting.

Base weighted CPI

Using base weights (Year 1 quantities), we will set the index in Year 1 to 100.

In Year 2, the index would be (using Year 1 weights):

$CPI_{Year\ 2} = CPI_{Year\ 1}$ × Total expenditure in Year 2 (Column 4) divided by Total expenditure in Year 1 (Column 3)

$CPI_{Year\ 2} = (100 \times \$42)/\$30 = 140$

Current weighted CPI

Using current weights (Year 2 quantities), we will again set the index in Year 1 to 100. In Year 2, the index would be (using Year 1 weights):

$CPI_{Year\ 2} = CPI_{Year\ 1}$ × Total expenditure in Year 2 (Column 3) divided by Total expenditure in Year 1 (Column 4)

$CPI_{Year\ 2} = (100 \times \$50)/\$36 = 138.9$

Thus we can see that it does make a difference which weighting approach we use.

Rate of growth of the CPI index

We have generated two CPI indexes (one base weighted and one current weighted) over two years, so we can calculate a measure of the overall movement in prices, and provide a measure of the change in the cost of living. The growth rate of the CPI measures the rate of inflation (if positive) or deflation (if negative), acknowledging that strictly inflation(deflation) is an ongoing, rather than a one-off, increase(decrease) in the price level.

We can write the percentage rate of inflation(deflation) as:

(4.6) $CPI_{Gt} = 100 \times [(CPI_t - CPI_{t-1})/CPI_{t-1}]$

where CPI_t denotes the index magnitude at time t and CPI_{Gt} denotes the change of the CPI from time $t - 1$ to time t (say, one year). So the rate of change can be expressed as 100 multiplied by the change in the index, divided by the initial value of the index.

It can be readily shown using the data in Table 4.2 that the respective rates of change for the base and current weighted price indexes between Year 1 and Year 2 are 40 per cent and 38.9 per cent.

You will appreciate that the current weighted index takes into account changes in prices and the quantities purchased following these price changes, whereas the base weighted approach considers price changes only and ignores the fact that people will change their expenditure patterns over time as relative prices change.

In practice, household expenditure patterns change and new goods and services are sold, so statisticians periodically revise the weights in the basket of goods and services in line with other information that they collect. They have complex methods to splice the new and the old indexes together. In the next subsection we explore the biases associated with using the CPI to accurately measure inflation.

Finally it should be recognised that there are other published price indexes, including those based on wholesale and retail prices. For example, the US Producer's Price Index is based on the wholesale prices of approximately 3,000 items, including raw materials and semi-finished goods.

Difficulties in using the CPI to accurately measure inflation

Measurement biases

There are many difficulties in using the CPI to get an accurate measure of inflation. For example, if consumers increase the percentage of purchases at 'discount' outlets, the CPI will overstate the actual rate of inflation experienced by the typical consumer. This is called the '*outlet substitution bias*' because the index does not adequately consider such shifts.

In addition, over time consumers will change the composition of the basket of consumer goods which they purchase. The composition of the basket used to calculate the price index is revised on an irregular basis, which results in a bias. Economists identify three different kinds of bias associated with changing baskets: substitution bias, quality change bias, and new product bias. In addition, there is growing recognition of a fourth kind of bias – the formula bias.

Substitution bias refers to the impact that changing relative prices would have on the composition of the basket. If for example, the price of tea rises relative to that of coffee, economic theory suggests that consumers will substitute coffee for tea. However, as the CPI basket might be changed only once per decade, it may be some time before the switch to coffee is reflected. The index will be calculated as if no substitution had occurred, leading to an overstatement of inflation due to a *substitution bias*.

Often when prices rise, this reflects increases of the quality of products (products might last longer or provide a higher level of services). In most cases, it is very difficult to calculate what portion of a price increase should be attributed to quality changes. The Bureau of Labor Statistics (BLS), for example, does not even attempt to calculate this for many products. Thus, inaccurate measures of quality change introduce a *quality change bias*, which would normally be expected to overestimate inflation, because it underestimates the quality changes that justify higher prices.

Thirdly, new products are introduced all the time. The BLS includes these in the basket only with long and variable lags, which introduces a *new product bias* into the CPI. In the case of some goods a considerable bias results. For example, many high technology consumer goods follow a price cycle that begins with very high prices for goods sold to high income classes, then prices fall rapidly as the goods are introduced to lower income classes, and then prices gradually rise again as the market matures. If the BLS introduces the goods into the basket only after prices have reached their minimum, the CPI will not capture the period during which prices fell rapidly, but will include the mature period over which prices rise. In recent years, this bias would be expected to be quite important due to the introduction of new consumer electronics (and the speed with which these can become obsolete).

There is yet another source of bias called the *formula bias*. This bias results because price data are collected on a disaggregated basis and then aggregated in a very complex manner that can introduce anomalies. For example, the calculation method used in recent years gives too much weight to items on sale; somewhat paradoxically, this generates formula-induced inflation as the items go off sale. The degree of this bias can increase with the frequency of rotation (of outlets included in the sample), because the bias results from short-run price variability and the use of a method that gives greater weight to lower-than-average prices.

Researchers had noticed that surveys of average prices actually paid by consumers showed rates of inflation well below the rates of inflation reported by the CPI for relatively disaggregated components of the consumer basket. While part of this could be attributed to the outlet substitution bias, most of it could not. Estimates of the formula bias run as high as six-tenths of a percentage point for owner occupied housing and one percentage point for apparel, an item often on sale.

The housing component

The housing component of the CPI is very large; in the US it is above 40 per cent of the index and during high inflation periods it contributes up to half of measured inflation. There are two ways to calculate the contributions of the housing sector to a price index: the flow of services approach or the homeowner or 'user' cost approach.

The method currently used in the US, imputed rental cost, is based on the flow of services approach and has been in place since 1983. Previously, the BLS had tried to calculate the user cost of housing, but it was believed that this method mixed the investment and consumption features of home ownership.

The largest portion of the housing component is 'shelter' services, which accounts for more than two-thirds of the housing component. Nearly three-quarters of the shelter component reflect owner-occupied costs, since most Americans own their own homes, and the rest are renters' costs. Most homeowners' costs are based on the owners' equivalent rent.

The BLS uses a survey of rental units to obtain data regarding changes of rental price. The results are adjusted through a weighted averaging process, and quality adjustments are made to deal with ageing and improvements. The imputation of renter's costs to be included in the CPI is therefore straightforward.

However, the method used for owners' equivalent rent (OER) is more complicated. Field agents ask owners for the rental price that the homeowner believes the house could secure. Agents may enter their own estimate if they believe the owner's estimate is unreasonable. These survey data are used to establish the base year imputed rent. Subsequent values of implicit rent for a given unit are obtained by applying the rate of increase of prices of rental units that are thought to be similar in certain respects (location, structure type, and quality) to the owner-occupied home.

There are situations in which this method of calculating 'inflation' of the housing component could lead to erroneous results (for example, where the statisticians impute high inflation when actual housing prices are falling). We need not go into that here. Instead, what we want to make clear is that the construction of an index is difficult, subject to controversial decisions, and to error.

Further, it is important to understand that the CPI comprises components that have 'imputed prices', prices formulated by statisticians rather than being obtained from markets. This is because statisticians want to obtain a measure of the cost of a relatively complete consumer basket that includes items that are not bought annually such as housing 'services' enjoyed by those who own their own homes.

There is a trade-off made between calculating a CPI that takes a 'hedonic' approach, by seeking to put a price on the 'enjoyment' received from the entire consumption basket, and one that attempts to focus on what is happening to the market prices of purchased items. The problem with the first is that statisticians must make many guestimates. The problem with the second is that it does not deal adequately with quality adjustments.

What all this means to the student is that you should take CPI measures of the inflation rate with the proverbial grain of salt! Especially at low measured rates of inflation, we cannot be sure if the prices of things people buy are truly rising, steady, or falling.

4.9 Measuring National Inequality

As discussed above, our measures of national output (GDP) and income (GNI) do not directly take account of the distribution of output and income. Economists typically use the **Gini coefficient** which is derived from a **Lorenz curve** as an index of income distribution. The Gini coefficient was developed by the Italian statistician and sociologist Corrado Gini in 1912. The Lorenz curve was devised by the American economist, Max Lorenz in 1905.

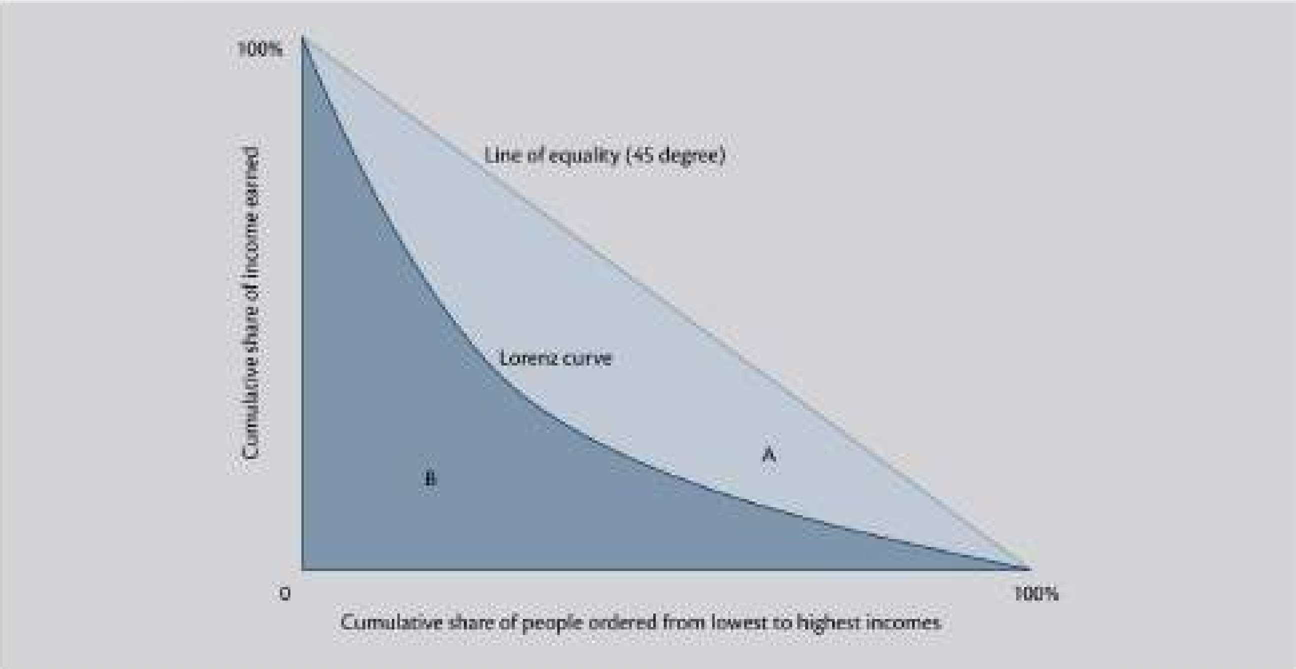
The Lorenz curve plots the share of total income received (vertical axis) by the lowest X per cent of income earners (horizontal axis) (see Figure 4.1). It is easy to see that in our example the distribution is not equal because as we move from the origin at the left end of the horizontal axis, the share of income going to those with the lowest income initially increases slowly. As we move to the higher-income people, the cumulative share of income increases more rapidly. The 45 degree line shows the case of perfect equality, so that 30 per cent of people have 30 per cent of total income; 60 per cent of people have 60 per cent of total income and so on.

We can calculate the Gini coefficient as a ratio, using the two areas, A and B in Figure 4.1:

(4.7) Gini coefficient $= A/(A + B)$

Different-shaped Lorenz curves can generate the same value of the Gini coefficient. In addition, there are different ways to measure income, for example before or after taxes, and before or after income transfers. The Gini coefficient can also be represented by an algebraic formula. There are also alternative indexes to the Gini coefficient.

Figure 4.1 The Lorenz curve



It is important to realise that different indexes exhibit different properties and the choice of which index to use should be made in light of the objectives associated with measuring inequality.

A Gini coefficient of zero means that income is perfectly equally distributed as the Lorenz curve coincides with the line of equality. Alternatively, a Gini coefficient of one means that income is perfectly unequally distributed: that is, one person has all the income.

Table 4.3 shows the Gini coefficients for all the nations that belong to the Organisation for Economic Co-operation and Development (OECD), for which comparable data is available for the years 2004 and 2012. The data are based on disposable income after taxes and transfers. The Gini coefficients mostly range between the values of 0.25 to 0.50. There is considerable diversity among these nations with respect to income inequality. Sweden had the least inequality in 2004, while Mexico had the highest inequality in both years. The USA has the highest inequality of the rich developed nations, while the Scandinavian countries tend to enjoy the lowest inequality. Note also that inequality increased in many nations between 2004 and 2012, while it declined in other nations (indicated by the + and – signs).

Table 4.3 Gini coefficients for several OECD nations, 2004 and 2012

Country	2004	2012	Change
Australia	0.315	0.324	+
Austria	0.269	0.276	+
Belgium	0.287	0.262	–
Czech Republic	0.269	0.252	–
Estonia	0.346	0.326	–
Finland	0.267	0.261	–
France	0.283	0.306	+
Germany	0.285	0.289	+
Greece	0.336	0.340	+
Iceland	0.262	0.252	–

Country	2004	2012	Change
Ireland	0.323	0.302	-
Italy	0.331	0.326	-
Luxembourg	0.263	0.299	+
Mexico	0.474	0.482	+
Norway	0.276	0.253	-
Poland	0.381	0.300	-
Portugal	0.383	0.341	-
Slovak Republic	0.266	0.249	-
Slovenia	0.247	0.251	+
Spain	0.332	0.335	+
Sweden	0.234	0.274	+
United Kingdom	0.331	0.351	+
United States	0.360	0.389	+

Source: Data from OECD Statistics. Gini coefficients for 2012 are based on a new definition of income that includes a more detailed breakdown of "current transfers received and paid by households as well as a revised definition of household income, including the value of goods produced for own consumption as an element of self-employed income" (OECD, 2016).

Conclusion

This chapter introduced the system of National Income and Product Accounts (NIPA), the framework adopted by countries to measure economic activity. Gross domestic product (GDP) and gross national income (GNI) were defined, and the components of each were examined. Some of the difficulties of measuring output were discussed. Alternative ways of measuring growth and inflation were presented, including most importantly the use of the consumer price index. Finally, the text discussed the Lorenz curve, and the Gini coefficient, which is used to measure national inequality.

References

Economist, The (2014) "Sex, Drugs and GDP: Italy's inclusion of illicit activities in its figures excites much interest", 31 May.

Landefeld, J.S. and Parker, R.P. (1997) "BEA's Chain Indexes, Time Series, and Measures of Long-Term Economic Growth", *Surveys of Current Business*, June, 58–68.

OECD (2016) *Income Distribution and Poverty: by country – Inequality*, OECD.Stat, Available at: http://stats.oecd.org/index.aspx?queryid=66670, accessed 22 August 2018.

Endnote

1. If the production of all final output is vertically integrated, so a single firm is responsible for all stages of the production for each good or service, then there is no intermediate consumption.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

5 LABOUR MARKET CONCEPTS AND MEASUREMENT

Chapter Outline

Learning Objectives

- Gain an understanding of the main features of the Labour Force Framework and the definitions of employment and unemployment.
- Recognise the deficiencies of measuring labour underutilisation by the official unemployment rate.
- Understand the relationships between labour market stocks and flows.
- Recognise the importance of unemployment duration in understanding labour market hysteresis.

5.1 Introduction

Chapter 3 provided an outline of the evolution of economic systems from tribal societies through slavery, feudalism and eventually to modern capitalism. An introduction to the concept of a (labour) market as a social construct with embedded power relations was developed to provide an intrinsic understanding of what happens when someone gets a job and receives a wage.

This chapter is largely devoted to definitional and measurement issues associated with modern labour markets. We outline the labour market framework, which incorporates definitions of the states of employment, unemployment and not in the labour force, which are stocks. We classify types of unemployment and argue that the rate of unemployment is an inadequate measure of labour underutilisation. We explain the relationships between our stock measures and the flows between the labour market states. We conclude with an exploration of the average duration of unemployment and its role in the process of labour market hysteresis.

5.2 Measurement

Many textbooks will state that macroeconomics is the study of the behaviour of employment, output and inflation in the economy as a whole. Thus, in addition to focusing on how real GDP, national income and prices are determined, macroeconomics also seeks to understand the dynamics of employment and relatedly, unemployment.

Further, a central idea in both microeconomics and macroeconomics is efficiency – getting the best out of the available resources. The concept is extremely loaded and is the focus of many disputes, some more arcane than others. At the macroeconomic level, the efficiency frontier is normally summarised in terms of full employment, which has long been a central focus of economic theory, notwithstanding the disputes that have emerged about what we mean by the term. However, despite doctrinal disputes about how to define full employment, most economists would agree that an economy cannot be efficient if it is not using all the resources available to it. In recent decades, the emergence of issues relating to climate change have focused our attention on what that resource limit is, or should be. In this chapter, we focus on the use of labour resources.

The quest for **full employment** was embodied in the policy frameworks and definitions of major institutions in most nations at the end of the Second World War. The challenge for each nation was how to turn its wartime economy, which had high rates of employment because of the prosecution of the war effort, into a peacetime economy, without sacrificing those high rates of labour utilisation.

In this section, we outline key concepts and consider issues relating to measurement. How do we know how much employment there is at any point in time? What is unemployment? Is it a measure of wasted labour resources or are there other issues that should be considered?

Labour force framework

The Labour Force Framework constitutes a set of definitions and conventions that allow national statisticians to collect data and produce statistics about the labour market. These statistics include employment, unemployment, economic inactivity, and underemployment, which can be combined with other survey data covering, for example, job vacancies, earnings, trade union membership, industrial disputes and productivity to provide a comprehensive picture of the way the labour market is performing.

The Labour Force Framework is a classification system, governed by a set of rules and categories. It forms the accepted foundation for cross-country comparisons of labour market data. The framework was conceived and made operational through the International Labour Organization (ILO) and its International Conference of Labour Statisticians (ICLS). These conferences and expert meetings developed the guidelines or norms for implementing the labour force framework and generating the national labour force data – see Box 5.1.

The Australian Bureau of Statistics (ABS) publication *Labour Statistics: Concepts, Sources and Methods* describes the international guidelines that have been agreed by the national statistical agencies. The guidelines outline the organising principles that define the Labour Force Framework. Thus, national statistical agencies (such as the Bureau of Labor Statistics in the US, the Office of National Statistics in Britain, the Federal Statistics Office in Germany) work within internationally agreed standards and criteria when publishing labour statistics.

The rules contained within the labour force framework have the following features:

- An activity principle, which is used to classify the population into one of the three basic categories, namely employed, unemployed and not in the labour force.
- A set of priority rules, which ensure that each person is classified into only one of the three categories.
- A short reference period to reflect the labour supply situation at a specified moment in time.

The priority rules are applied to ensure that labour force activities take precedence over non-labour force activities, and working or having a job (employment) takes precedence over looking for work (unemployment). Also,

BOX 5.1 THE COLLECTION AND PUBLICATION OF LABOUR MARKET STATISTICS

At the end of the First World War, the ILO was established to set minimum labour standards. Each year, there is an International Labour Conference which makes decisions that determine what are called the International Labour Conventions and Recommendations.

The Labour Statistics Convention (No. 160) was adopted at the 71st International Labour Conference in 1985 and modernised the previous convention that was agreed in 1938.

Article 1 of the 1985 Convention requires all member states of the ILO (including Australia and the US) to:

regularly collect, compile and publish basic labour statistics, which shall be progressively expanded in accordance with its resources to cover the following subjects:

a. *economically active population, employment, where relevant unemployment, and where possible visible underemployment;*
b. *structure and distribution of the economically active population, for detailed analysis and to serve as benchmark data;*
c. *average earnings and hours of work (hours actually worked or hours paid for) and, where appropriate, time rates of wages and normal hours of work;*
d. *wage structure and distribution;*
e. *labour cost;*
f. *consumer price indices;*
g. *household expenditure or, where appropriate, family expenditure and, where possible, household income or, where appropriate, family income;*
h. *occupational injuries and, as far as possible, occupational diseases; and*
i. *industrial disputes.* (ILO, 1985)

The ILO also publishes very detailed technical guidelines about how these statistics should be collected and disseminated via one of its technical committees, the International Conference of Labour Statisticians (ICLS). This committee meets about every five years and is comprised of government officials "mostly appointed from ministries responsible for labour and national statistical offices" (ILO, n.d.), and representatives from employers' and workers' organisations.

The ICLS agree on resolutions, which then determine the way in which the national statistical offices collect and publish data. While the national statistical agencies have some discretion as to how they prepare labour statistics, in general there is widespread uniformity across agencies.

Labour statistics are often drawn into political controversies and critics have been known to accuse the government of manipulating the official data for political purposes. But once you understand the process that governs the structure of the labour force statistical collection and the definitions outlined in the ICLS resolutions, it is hard to believe that argument.

This is not to say that there is not a lot of debate about what the official labour statistics measure and whether they can be improved, but it is important to understand how they are collected.

as with most statistical measurements of activity, employment in the informal sectors or the black market economy is outside the scope of activity measures.

There is a long-standing concept of 'gainful work' which shapes these priorities, but this has proven controversial. Gainful work is typically seen as work for profit or pay. One can work in government or in the non-profit sector where a payment is received, but unpaid work of any kind is not included.

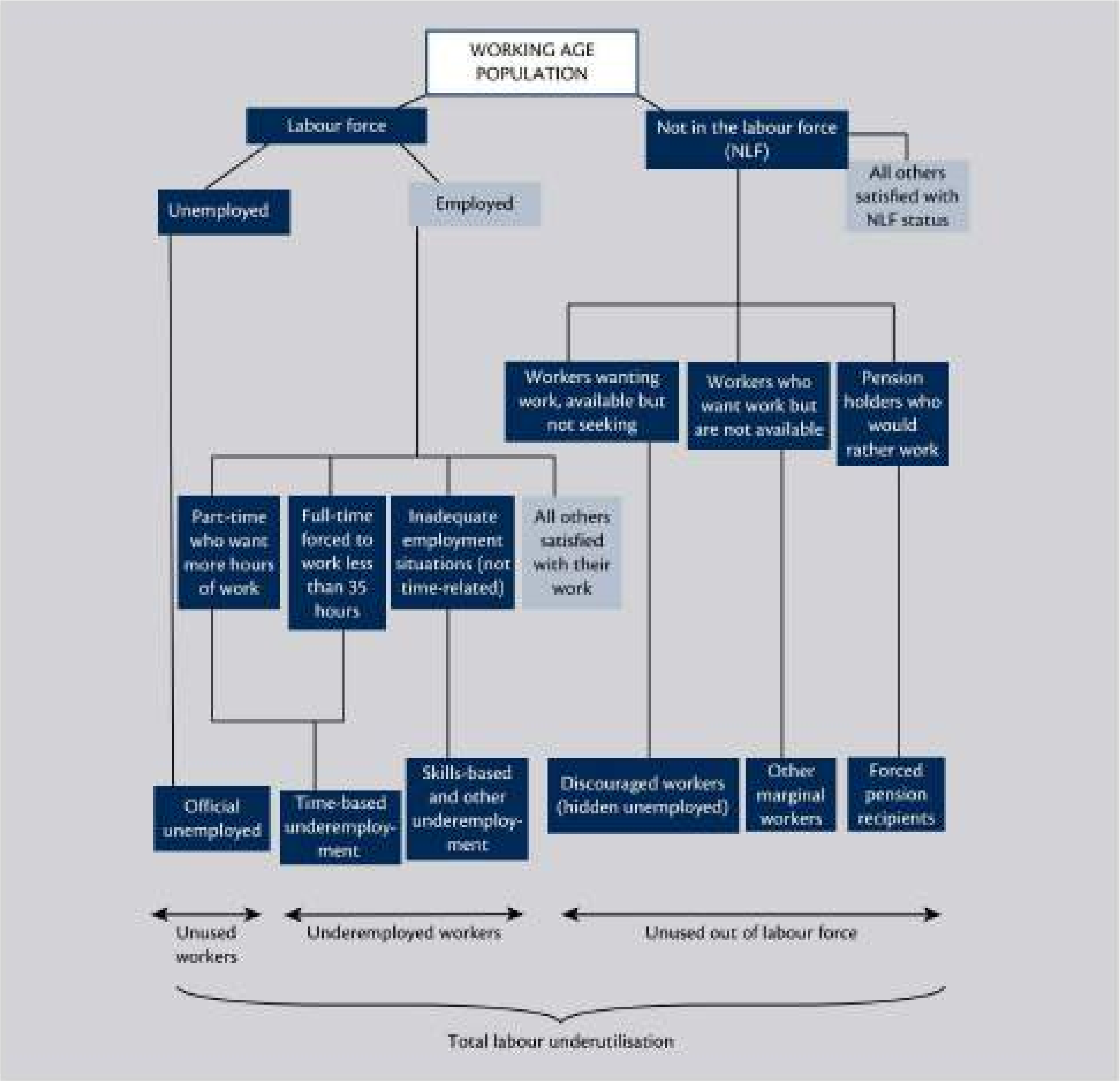
Thus, a person who does ironing for a commercial laundry is defined as pursuing gainful work, whereas if the same person is ironing only for their family they are considered inactive. Clearly, with economic and non-economic roles being biased along gender lines, this distinction leads to an undervaluation of a substantial portion of work performed by women, as we noted in Chapter 4.

Another example of the precedence of paid over unpaid activities is that persons who maintain a house on an unpaid basis are classified as not in the labour force, while those who receive pay for this activity are in the labour force and employed. Similarly, persons who undertake unpaid voluntary work are not in the labour force, even though their activities may be very similar to those undertaken by the employed.

Figure 5.1 summarises the Labour Force Framework as it applies in Australia, but this structure is common across all nations. National statistical agencies conduct a **Labour Force Survey** (LFS) on a regular basis, usually monthly, to collect data using the concepts and definitions provided in the Labour Force Framework.

The **working age population** (WAP) typically refers to all citizens above 15 years of age. In several countries, the lower age threshold is 16. In the past, the age span was from 15 years old to retirement age, usually around

Figure 5.1 The labour force framework



65 years. However, as social changes have seen age discrimination laws come into force in many countries, the upper age limit has been accordingly abandoned in several nations. Also, the age at which retirees can access a government pension has been increased in a number of countries as part of wider austerity measures.

The WAP is broken down into the **Labour force** (the 'active' component) and Not in the labour force (the 'inactive' component). A worker is considered to be active if they are **employed** or **unemployed**.

The proportion of the adult population who comprise the labour force is governed by the **labour force participation rate**, which is defined by the ILO as:

> **the ratio of the labour force to the WAP, expressed as a percentage.**

We will consider the cyclical behaviour of the participation rate later in the chapter.

The ILO defines a person as being employed if, during the reference period that may be as short as one week or even one day, they satisfied one of the following:

a. performed some work for wage or salary in cash or in kind,
b. had a formal attachment to their job but were temporarily not at work during the reference period,
c. performed some work for profit or family gain in cash or in kind, or
d. were with an enterprise such as a business, farm or service but who were temporarily not at work during the reference period for any specific reason.
(ILO, 1988: 47).

What constitutes 'some work' is unclear and controversial. In Australia and the USA, for example, a person who works one or more hours a week for pay is considered employed, which makes the demarcation line between employed and unemployed very thin – a single hour of paid work.

Within the employment category further subcategories exist, which we will consider later. Most importantly, significant numbers of employed workers might be classified as being **underemployed**, if they are not able to work as many hours as they desire, because there is insufficient aggregate spending in the economy at that point in time.

What constitutes unemployment? According to ILO concepts, a person is unemployed if they are over a particular age, they do not have work, but they are currently available for work and are **actively seeking work**.

Unemployment is therefore defined as the difference between employment and the economically active population (civilian labour force).

Two derivative measures capture a lot of public attention. First, the **unemployment rate** is defined by the ILO as:

> **the number of unemployed persons as a percentage of the civilian labour force.**

To see this calculation in practice, the US unemployment rate in November 2016 was 4.6 per cent. This was derived from a labour force estimate of 159.486 million and total estimated unemployment of 7.4 million.

Second, statisticians publish the **employment-population ratio**, which is:

> **the proportion of an economy's working age population that is employed.**

To see this calculation in practice, the data published by the Australian Bureau of Statistics reveals that in November 2016, total employment was 11,973.2 thousand and the working age population was 19,642.7 thousand. This gives an employment-population ratio of 61 per cent.

Note that the denominators of these two ratios, the unemployment rate and the employment-population ratio, are different. The unemployment rate uses the labour force while the employment-population ratio uses the working age population, the former being a subset of the latter.

We will see why this difference matters when we consider the way the labour market adjusts over the economic cycle, and how this impacts on our interpretation of the state of the economy.

The unemployment rate is what economists refer to as a stock measure. It is defined as a ratio of two stocks: the number of unemployed (numerator) and the labour force (denominator). The stock measure of the unemployment rate is compiled by the national statistician at a point in time, usually monthly.

The ILO states that persons **marginally attached to the labour force** are those who are not economically active under the standard definitions of employment and unemployment, so they are out of the labour force, but who, following a relaxation in one of the standard definitions of employment or unemployment, would be reclassified as economically active. Thus, for example, a relaxation of the criterion used to define availability for work (say from within one week to within four weeks) will increase the numbers of people classified as unemployed. This leads to the likelihood of volatility in the series and thus there can be endless argument about the limits applied to define the core series.

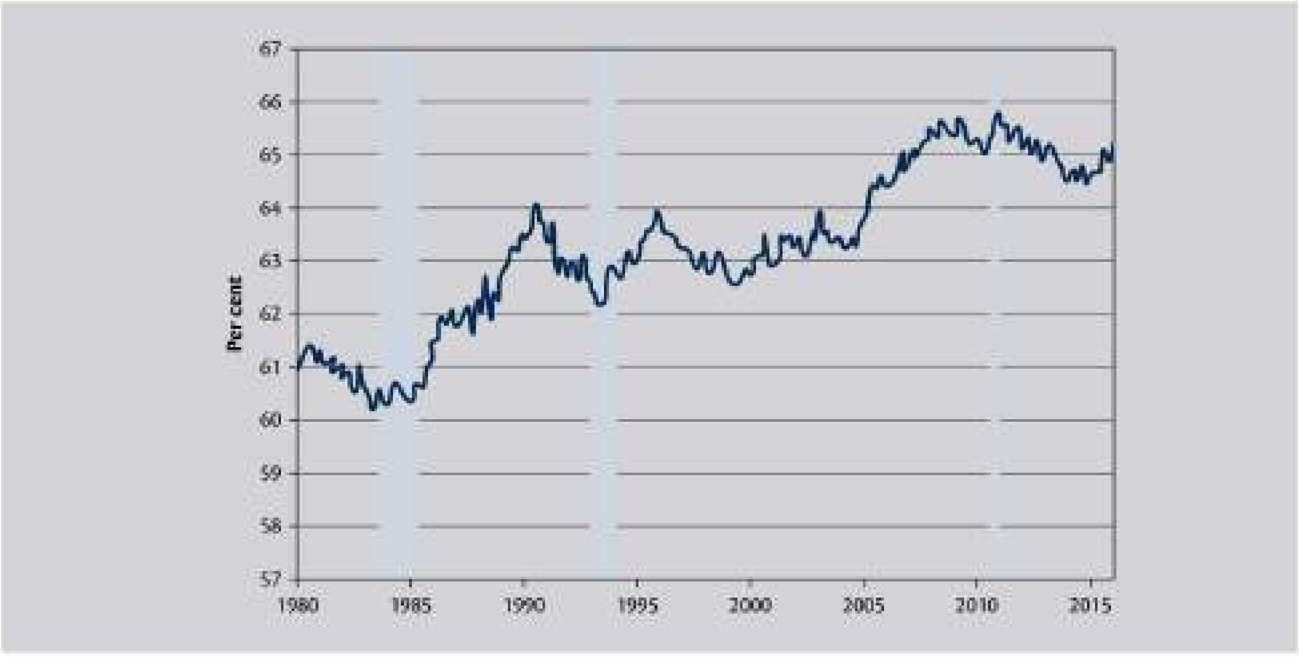
Impact of the business cycle on the labour force participation rate

The working age population is the population aged above the minimum working age, which is usually set at 15. You have learnt that the proportion of the working age population that offers themselves for work is called the labour force participation rate. A change in the participation rate leads to a change in the size of the labour force for any given WAP level.

The labour force participation rate is a procyclical variable; it rises in good economic times and falls when job opportunities are scarce. This means that in bad times there are likely to be some workers who would be willing to take jobs if they were offered, but who have stopped looking for work and are classified by the national statistician as being not in the labour force. The workers who are discouraged from job search by the apparent lack of job opportunities, are classified as being **hidden unemployed**. From the perspective of availability, these workers are no different to the officially recorded unemployed. If a job offer was made to them they would take it promptly. This suggests that in bad times, the official unemployment rate understates the 'true' underlying unemployment rate in the economy due to the lower rate of labour force participation.

Figure 5.2 shows the labour force participation rate for Australia from January 1980 to December 2015. The grey columns denote periods of slowing economic activity. The pattern has two features that are common in the participation rate of most nations. First, it is clear there has been an overall upward trend in participation over

Figure 5.2 Labour force participation rate, Australia, 1980 to 2015, per cent



Source: Authors' own. Data from Australian Bureau of Statistics, Labour Force Survey.

this time, largely the result of the increased involvement of married women in the labour market. Second, there are distinct cyclical episodes coinciding with fluctuations in real GDP growth.

For example, in the early 1990s there was a severe recession in Australia, which precipitated a major decline in the participation rate. Participation then grew in the early 2000s with the growth in employment opportunities.

With the onset of the Global Financial Crisis in early 2007 and slowing employment growth, the participation rate fell because job opportunities became scarcer. In late 2008, the Australian government reacted to the crisis by introducing two large fiscal stimulus packages, which promoted growth and an improvement in labour market conditions. The unemployment rate fell until early 2011, but then steadily increased to over six per cent in early 2015, which coincided with the fall in the participation rate.

5.3 Categories of Unemployment

Economists have long used taxonomies to organise their thoughts about unemployment. We shall address two of the most popular. These taxonomies can cut across each other and no single category is better than the others. The categorisation system used depends on the purpose of the analysis. In general, economists have married these categorisation frameworks into broader theoretical discussions which seek to explain why unemployment arises, whether it is a problem or not, and what can be done about it via policy interventions should we consider it to be a problem.

The most frequently used typology for unemployment distinguishes between frictional, seasonal, structural, and cyclical (demand deficient) unemployment.

Frictional unemployment – recognises that the labour market is in a constant state of flux. Jobs are continually being created and destroyed, so workers who have been laid off or have quit are moving between jobs, while firms seek workers for newly created jobs or to fill existing jobs if the previous incumbents have left.

Further, new entrants into the labour force seek work while retirees leave jobs. Frictional unemployment arises because the matching of these demand and supply flows is not instantaneous. It takes time for workers and employers to gather relevant information and for the former to secure employment. Frictional unemployment is a short-term phenomenon and part of the normal functioning of the labour market. This category would be expected to comprise around one to two per cent of the labour force.

Seasonal unemployment – arises when certain occupational skill groups and industry sectors experience predictable fluctuations of a systematic (seasonal) nature over the course of the year. For example, in certain regions, workers who are engaged in harvesting agricultural crops will experience seasonal unemployment when they move between crops and localities. This category is small in magnitude when assessed on a macroeconomic scale. It is also difficult to distinguish from frictional unemployment.

Structural unemployment – is said to arise when there are enough jobs available overall to match the total pool of unemployment but there are mismatches between the skills demanded and the skills supplied and/or between the location of the jobs available and the location of the unemployment. This category of unemployment is often discussed in the context of industrial restructuring (for example, the decline of the manufacturing sector or deindustrialisation). Changes in the composition of industry employment create job losses in declining sectors and new job opportunities in emerging sectors. Further, given that industry employment is not spread evenly across regional space, the decline of a major firm in one region will have significant implications for the local labour market.

Changes in technology also have structural impacts in the sense that new skills become relevant, while old skills cease to be in demand by firms.

These disruptions to the pattern of employment take time to resolve. The relocation and retraining of workers displaced by structural change is sometimes a lengthy process. It is the changing pattern of required skills, the changing location of jobs and the extended time taken to resolve the resulting demand and supply imbalances that distinguish the concept of structural unemployment from frictional unemployment.

However, there are two important qualifications to the normal conceptualisation of structural unemployment which are not often considered in the mainstream textbooks.

First, the idea of a skills shortage is a relative concept. Unsurprisingly, analyses of skills shortages by industry and governments invariably take the perspective of business and profitability, with an emphasis on the containment of labour costs both in terms of wages and conditions. Whenever possible, any costs associated with developing the skills that firms require in their workers are externalised.

Within this context the notion of structural unemployment arising from 'skills mismatch' can be understood as implying an unwillingness of firms to offer jobs, with attached training opportunities, to unemployed workers whom they deem to fall short of their ideal profile. When the labour market is tight, this selectiveness is more costly and firms are more likely to lower their hiring standards and even package training opportunities with job offers.

However, when labour underutilisation is high, firms can easily increase their hiring standards, that is, broaden the desired characteristics demanded from workers. The training dynamism driven by labour shortages is then absent. In this case, we observe, in a static sense, 'skill mismatches' as symptoms of a 'low pressure' economy.

Thus, hiring standards and the willingness of firms to provide training opportunities when making job offers vary with economic activity. This means that structural unemployment is difficult to distinguish from demand deficient unemployment, which is related to a lack of aggregate demand in the economy and, like these 'skills mismatch' problems, is cyclical in nature.

Hence, there are significant overlaps between these categories, which reduce their capacity to provide a definitive decomposition of total unemployment.

Cyclical (demand deficient) unemployment – arises when there is a shortage of jobs overall relative to the willing supply of labour resources (persons and hours) at the current wage levels. This category is termed demand deficient unemployment because it relates to a deficiency in aggregate demand. Unemployment thus varies over the economic cycle, rising when aggregate spending falls below the level needed to fully employ the available workforce and falling when aggregate spending moves closer to the level needed to fully employ the available supply of labour.

Cyclical unemployment is also known as mass unemployment and arises when the macroeconomic system fails to generate enough jobs to match the preferences of the available workforce. It is also related to the concept of an **output gap**, which measures the percentage deviation of real GDP from the potential production levels at any point in time.

During an economic downturn (which may become a recession), cyclical unemployment will be the dominant proportion of measured unemployment. When economic activity improves because of increased aggregate demand, cyclical unemployment falls.

In Chapter 19 we will see that the economic and social costs of unemployment (associated with output gaps) are enormous, which makes the elimination of cyclical unemployment a policy imperative. The solution to cyclical unemployment is thus to increase the growth rate of aggregate demand to close any output gaps.

5.4 Broader Measures of Labour Underutilisation

Figure 5.1 summarised the Labour Force Framework for Australia, and operating through the ILO and its ICLS. All national statistical agencies have broadly similar structures for collecting data about the labour market.

We focus on unemployment as an indicator of labour market performance because it signifies a waste of productive resources, quite apart from the individual and social costs that accompany it. However, unemployment is a narrow measure of labour underutilisation.

Labour underutilisation arises for a number of different reasons that can be subdivided into two broad functional categories:

- **A category involving unemployment or its near equivalent** – In this group, we include the official unemployed under the ILO criteria plus those classified as being not in the labour force due to failing to search for employment (discouraged workers), those who are unavailable to start work (other marginal workers), and more broadly still, those who take disability and other pensions as an alternative to unemployment (forced pension recipients). These workers share the characteristic that they are jobless and desire work if there were available vacancies. However, they fail to satisfy all the criteria for being defined as unemployed. In particular, they are not actively seeking work.

- **A category that involves suboptimal employment relations** – Workers in this category satisfy the ILO criteria for being classified as employed but suffer time-related underemployment, which is typically associated with part-time work, or inadequate employment situations (working below their skill level).

We have already considered the concept of hidden unemployment, which is a near-equivalent state to unemployment. In this section, we focus on underemployment, which has become an increasingly significant problem for most nations, especially following the 2008 Global Financial Crisis.

We have seen that within the Labour Force Framework, a person of working age is considered employed if they have worked a minimum number of hours in the reference week for pay, typically at least one hour per week. (The hours requirement differs across countries, but it is always a low threshold.) Otherwise, they are classified as unemployed or not in the labour force, depending on how they meet the activity criterion.

However, a person working part time might desire to work more hours but the state of the economy precludes this. This person is classified as **underemployed**. In this case, the underemployment is time related, referring to employed workers who are constrained by the demand side of the labour market to work fewer hours than they desire. Workers in this state are sometimes said to be in visible underemployment.

The other category of underemployment is termed 'skill related' and refers to workers who undertake jobs which have skill demands below their qualifications. Clearly, if society invests resources in education, then the skills developed should be used appropriately to maximise return on that investment.

The concept of an inadequate employment situation (particularly, skill-related underemployment) is very difficult to quantify and this has led to a paucity of data being available to measure it. However, national statisticians have developed sophisticated measures of time-related underemployment or visible underemployment.

In conceptual terms, at different moments of the reference time period, an underemployed worker can be considered both employed and unemployed, even though the worker is officially classified as employed throughout the whole period. Underemployed workers vary in their desires for extra hours of work.

An economy with many part-time workers who desire but cannot find full-time work is less efficient than an economy which reflects workers' preferences for work hours being satisfied. In this regard, involuntary part-time workers share characteristics with the unemployed. Table 5.1 shows the evolution of underemployment in a selection of OECD nations since 1990, ranked highest to lowest as at 2015.

Time-related underemployment, like unemployment, arises from a deficiency in aggregate demand. Unemployment manifests as a lack of available jobs, whereas the presence of underemployment indicates that the demand constraint rations the hours of work that are offered by firms. In both cases, willing labour resources are wasted.

The two concepts of underemployment are also related. The rising incidence of underemployment in many countries since the 1990s has been associated with a rising casualisation of the workforce as governments have tilted the industrial relations playing field towards employers and reduced workplace protections, including restrictions on the use of non-standard hours of work.

As a result, the quality of employment has fallen for many workers. Often part-time workers do not have access to the same benefits and protection that full-time workers enjoy. Some employers reduce hours to evade labour regulations that apply to full-time workers. This trend has also coincided with the growth of the service sector. In many nations, this growth has been concentrated in lower-skilled, less stable jobs. Underemployment is common in these sectors.

We will see in Chapter 19 that a meaningful definition of full employment should include zero underemployment. A worker cannot be considered fully employed if they are enduring underemployment.

Table 5.1 OECD underemployment, per cent of labour force, 1990 to 2015

	1990	2000	2005	2010	2015
Italy	1.6	4.6	8.1	12.1	16.9
Spain	1.1	3.0	6.7	9.1	11.7
Australia	3.8	8.2	9.4	9.6	11.1
Sweden	2.2	5.0	4.0	10.4	8.4
Japan	1.1	–	7.0	8.5	6.7
Canada	3.1	6.2	6.3	6.8	6.3
New Zealand	4.1	8.3	5.2	5.8	6.1
Greece	1.1	2.6	3.7	4.2	6.1
Portugal	1.3	4.1	4.0	4.5	5.9
OECD countries	1.7	3.0	4.3	5.2	5.4
Netherlands	5.4	2.1	2.4	3.4	5.3
Germany	0.7	3.9	6.9	7.7	5.3
Finland	–	5.1	4.4	4.6	5.1
United Kingdom	1.3	3.0	2.5	4.3	4.9
Austria	–	2.9	4.0	4.2	4.8
Luxembourg	0.4	1.6	3.5	2.6	4.3
Belgium	3	6.6	5.6	4.0	3.3
Norway	3.2	2.5	3.5	2.3	2.1
United States	–	0.9	1.2	2.0	1.8

Source: Data from OECD.Stat.

5.5 Flow Measures of Unemployment

The stock measure of each state indicates the level at a point in time. However, in each period there are large numbers of workers that flow between states: employment (E), unemployment (U) and not in the labour force (N). National statisticians measure these flows in their regular labour force surveys. The various stocks and flows are denoted as follows. Single letters denote stocks and two letters denote the flows between the stocks:

- E employment, with subscript t denoting the current period, and $t+1$ the next period
- U unemployment
- N not in the labour force
- EE flow from employment to employment (that is, the number of people who were employed last period and who remain employed this period)
- UU flow of unemployment to unemployment (that is, the number of people who were unemployed last period and who remain unemployed this period)
- NN flow of those not in the labour force last period and who remain in that state this period
- EU flow from employment to unemployment
- EN flow from employment to not in the labour force
- UE flow from unemployment to employment
- UN flow from unemployment to not in the labour force
- NE flow from not in the labour force to employment
- NU flow from not in the labour force to unemployment

Table 5.2 Labour market flows matrix

	Status in Period 1		
Status in Period 0	**Employed**	**Unemployed**	**Not in the labour force**
Employed	EE	EU	EN
Unemployed	UE	UU	UN
Not in the labour force	NE	NU	NN

Table 5.2 provides a schematic representation of the flows that can occur between the three labour force framework states.

To give you some idea of the magnitude of these flows between any given months, Table 5.3 summarises the flows for the US labour market for the period between December 2015 and January 2016.

The data in Table 5.3 show us that total US employment in December 2015 was 149.678 million, total unemployment was 7.541 million and the number of persons who were counted as being not in the labour force was 94.495 million. The sum of these stocks is equal to the WAP (the population above the age of 16 years) of 251.714 million.

The flows data show that between December 2015 and January 2016, 1.594 million workers who were unemployed in December 2015 moved into employment (UE) by January 2016. Similarly, 2.105 million workers counted as being employed in December 2015 had moved into the unemployment pool (EU) in January 2016.

In terms of flows into the category of Not in the labour force, 4.818 million workers who were counted as being employed in December 2015 exited the labour force (EN) in January 2016, and 1.859 million workers who were counted as being unemployed in December 2015 left the labour force (UN) in January 2016.

Flowing into the labour market, were 4.444 million new entrants who became employed (NE) and 2.099 million new entrants who ended up in unemployment (NU) in January 2016.

The final column and row show the levels of Employment, Unemployment and Not in the labour force, corresponding to December 2015 and January 2016, respectively. They are obtained from the row sums and the column sums. Employment fell over the month, which reflects the seasonal nature of employment in the winter, when some outdoor work cannot be conducted.

It is important to recognise that Table 5.3 tracks the labour force status of the 252.397 million US citizens who were part of the WAP in both December 2015 and January 2016. It does not include individuals who joined the WAP in January 2016, due to age or moving to the USA, and those who left the WAP, due to death or departure from the USA.

We can also calculate the total inflows and outflows from the three labour force states between any two periods of interest. Table 5.4 shows these calculations based on the data in Table 5.3. The total inflow into employment is measured by the sum, NE + UE, and for the period shown equalled 6.038 million, whereas the total outflow from employment, measured by the sum, EU + EN was 6.923 million. The net flow was thus negative and equalled 0.885 million workers. This confirms that employment between December 2015 and January 2016 fell.

The total inflow into unemployment is measured by the sum, EU + NU and for the period shown equalled 4.204 million, whereas the total outflow from unemployment, measured by the sum, UE + UN was 3.453 million.

Table 5.3 Gross flows in the US labour market, December 2015 to January 2016, millions

	Status Current Period			
Status last period	**Employed**	**Unemployed**	**Not in labour force**	**Stocks last period**
Employed	142.755	2.105	4.818	149.678
Unemployed	1.594	4.088	1.859	7.541
Not in labour force	4.444	2.099	87.952	94.495
Stocks current period	148.793	8.292	94.629	251.714

Source: US Bureau of Labor Statistics.

Table 5.4 Total inflow and outflow from labour force states, US, December 2015 to January 2016, millions

Labour market state	Total inflow	Total outflow
Employment	UE + NE = 6.038	EU + EN = 6.923
Unemployment	EU + NU = 4.204	UE + UN = 3.453
Not in the labour force	EN + UN = 6.677	NE + NU = 6.543

Source: US Bureau of Labor Statistics.

The net flow was thus positive (meaning that unemployment rose over the period) and was equal to 0.751 million workers.

Finally, the total exits from the labour force (into Not in the labour force) is measured by the sum, EN + UN and for the period shown equalled 6.677 million whereas the total new entrants into the labour force, measured by the sum, NE + NU was 6.543 million. The net flow was thus negative and equal to 0.134 million workers.

Labour market stocks and flows

We can understand changes in the stock measures associated with the labour market states from one period to the next by considering the net flows between two periods.

Total employment at any point in time (E_t) is given by Equation 5.1, which is strictly an identity, since an adult member of the population can only be in one of three labour market states at a point in time:

$$(5.1) \qquad E_t = E_{t-1} + UE_t + NE_t - EU_t - EN_t$$

In terms of the actual flows in the US labour market between December 2015 and January 2016 summarised in Tables 5.3 and 5.4, Equation 5.1 is evaluated as (in millions):

$$(5.1a) \qquad E_t = 149.678 + 1.594 + 4.444 - 2.105 - 4.818 = 148.793$$

The change in employment in any period, ΔE is the total inflows minus the total outflows:

$$(5.2) \qquad \Delta E = E_t - E_{t-1} = UE_t + NE_t - EU_t - EN_t$$

Total unemployment at any point in time (U_t) is given by:

$$(5.3) \qquad U_t = U_{t-1} + EU_t + NU_t - UE_t - UN_t$$

Equation 5.3 is evaluated as (in millions):

$$(5.3a) \qquad U_t = 7.541 + 2.105 + 2.099 - 1.594 - 1.859 = 8.292$$

Thus, the change in unemployment in any period, ΔU is the total inflows minus the total outflows:

$$(5.4) \qquad \Delta U = U_t - U_{t-1} = EU_t + NU_t - UE_t - UN_t$$

We can use the data from Table 5.3 to calculate so-called transition probabilities, which are the probabilities that transitions (changes of state) occur. These are obtained by dividing the elements of a row of the 3 × 3 flow matrix by the corresponding row sum.

In Table 5.5, the interpretation of (say) 0.03 in the first row is that there is a probability of 0.03 that an individual who was employed in December 2015 left the labour force over the following month. Allowing for rounding errors the rows each sum to unity.

We typically express these probabilities in percentage terms, such that the probability of a worker who was unemployed in December 2015 of becoming unemployed in January 2016 was 21 per cent. This likelihood was

Table 5.5 Labour market state transition probabilities, US, December 2015 to January 2016

	Status current period		
Status last Period	**Employed**	**Unemployed**	**Not in labour force**
Employed	0.95	0.01	0.03
Unemployed	0.21	0.54	0.25
Not in labour force	0.05	0.02	0.93

lower than the probability of 25 per cent that a person would transition from unemployment to being classified as not in the labour force.

Economists thus consider the labour market to be very dynamic and the extent of this dynamism is measured by the gross flows between the three labour market states, which is also revealed by the transition probabilities between the three states, as opposed to the probabilities on the main diagonal of the transition matrix (EE, UU, and NN), which measure the probability of remaining in the same labour market state.

Further, these flows are highly cyclical. For example, in a recession the flow EU increases while the flow UE declines. Workers also drop out of the labour force in greater numbers during a recession, so that labour force participation drops, and more new labour market entrants enter the state of unemployment, rather than employment.

5.6 Duration of Unemployment

As noted, the unemployment rate is a narrow measure of labour market performance. Another dimension of labour underutilisation, which it does not capture, is the duration of unemployment. As the discussion of flows indicated, the labour market is a very dynamic part of the economy with large flows between the labour force states occurring on a weekly basis. The magnitude of these flows is however highly cyclical and net flows into unemployment are larger during a recession than at other times.

It is therefore important to consider the average duration of unemployment as part of our assessment of the state of the labour market.

The Australian Bureau of Statistics' *Labour Statistics: Concepts, Sources and Methods* provides the following definition:

> *Duration of unemployment is defined as the elapsed period to the end of the reference week since the time a currently unemployed person began looking for work, or since a person last worked for two weeks or more, whichever is the shorter. Brief periods of work (of less than two weeks) since the person began looking for work are disregarded. (ABS, 2007)*

This conceptualisation is representative across nations even if there are some country-by-country variations in how the Labour Force Survey is collected.

The duration of unemployment influences the way we assess the distributional impacts of a recession. If for example, individuals who become unemployed only endure short spells of unemployment, that is, average duration in weeks is low, then the impact on their income flow and accumulated saving will be lower than if the spells of unemployment are longer. A drawn-out recession typically has the effect of wiping out any savings that an unemployed person may have accumulated.

For a given unemployment rate, an economy might be characterised by a predominance of short spells of unemployment (many people flowing in and out of the unemployment pool) or at the other extreme, fewer people enduring long spells of unemployment (low inflows into and low outflows out of the unemployment pool).

While any unemployment above some irreducible minimum rate is problematic, clearly the situation where individuals experience unequal durations of unemployment is more costly, both from an individual and an economic perspective.

As an example, assume an economy has a labour force of 100 persons and is enduring an unemployment rate of eight per cent. This might occur if eight individuals had become unemployed at the beginning of the month but who will find work in the following month. Next month, eight different individuals become unemployed. Thus, any one individual has a duration of unemployment of one month.

On the other hand, the same economy might have the same eight individuals enduring unemployment month after month and still maintaining an unemployment rate of eight per cent. Thus eight individuals had 12 months of continuous unemployment, whereas the remaining 92 individuals remained employed for the whole year. The total impact of unemployment upon these two hypothetical economies will obviously be very different.

The duration of unemployment displays distinct cyclical patterns. As economic activity starts to slow and the economy enters a recession, there are large flows into the unemployment pool and so short-term unemployment surges. As a result, in the early months of a recession, the overall pool of unemployment is more weighted to individuals with short duration spells of unemployment. As the recession endures and the net inflows into unemployment remain positive but start to decrease, more workers move into longer duration categories of unemployment and long-term unemployment increases. The average duration of unemployment starts to rise more sharply at this stage. The longer the recession, the higher will be the long-term unemployment rate.

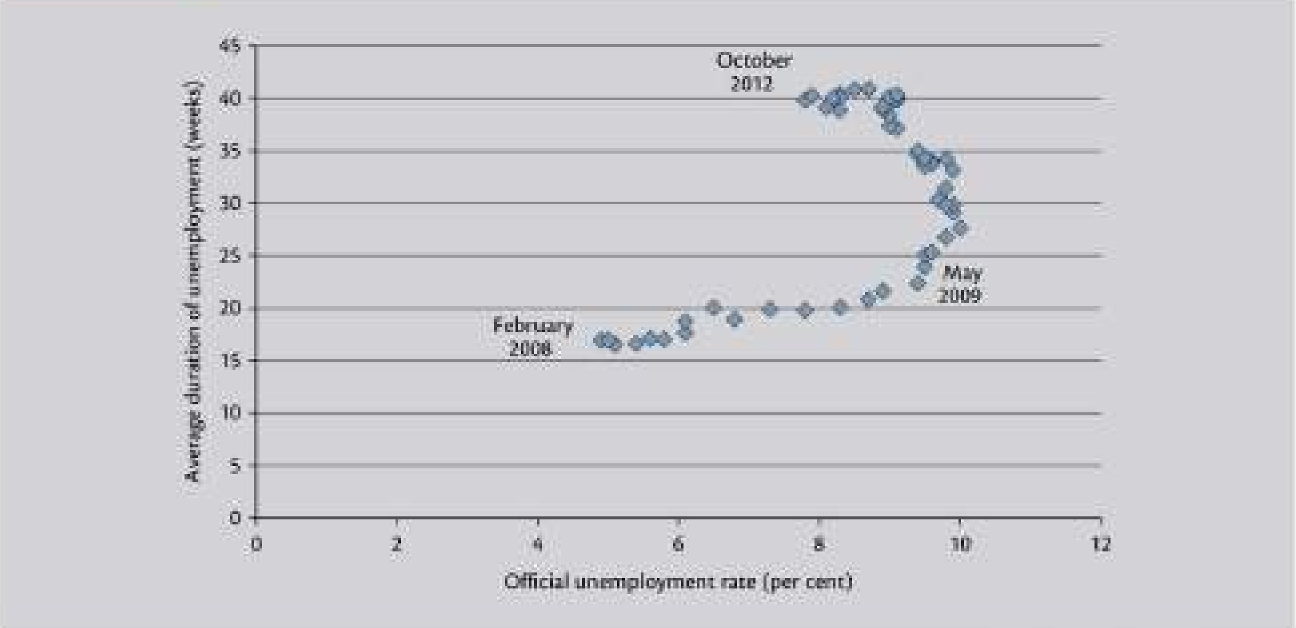
This pattern endures even as the economy recovers. As the flows into unemployment start to fall, the pool of unemployment is now more heavily weighted by individuals with longer spells of unemployment. As a result, the average duration of unemployment continues to increase even though the unemployment rate might start falling.

The problem is that in the early stages of the recovery, employment growth has to be strong enough to absorb the new entrants into the labour force (that is, to keep pace with the underlying population growth) and start eating into the huge pool of unemployed. There is evidence, which we discuss later, to suggest that in the early stages of the recovery, firms prefer to employ workers who have only endured short spells of unemployment. In other words, the longer a person has been unemployed, the lower will be the probability of them gaining work (see Section 5.7).

There are international variations in the official definition of long-term unemployment. For example, in the US, it is defined as a duration of unemployment of six months or more, whereas in Australia and the UK it is defined as a duration of 12 months or more.

Figure 5.3 illustrates the way in which the average duration of unemployment behaves during a downturn and early stages of recovery.

Figure 5.3 Unemployment rate and average duration of unemployment (weeks), US, February 2008 to October 2012



Source: US Bureau of Labor Statistics.

In February 2008, the official US unemployment rate was 4.9 per cent and the average duration of unemployment was 16.9 weeks.

In the first 12 months of the downturn, the unemployment rate increased by 2.9 percentage points and the average duration of unemployment rose by 2.9 weeks.

However in the second year of the downturn, the unemployment rate increased by 1.4 percentage points but the average duration of unemployment rose by 10.2 weeks. Even as the unemployment rate started to decrease in the third year of the crisis (by 0.7 percentage points), the average duration of unemployment increased by a further 7.3 weeks.

5.7 Hysteresis

One of the reasons we worry about situations where the duration of unemployment is high for extended periods relates to the concept of path dependence or **hysteresis.**

Hysteresis is a term drawn from physics and is defined by the *Oxford Dictionary* as:

> **The phenomenon in which the value of a physical property lags behind changes in the effect causing it, as for instance when magnetic induction lags behind the magnetizing force.**

In economics, we sometimes say that where we are today reflects where we have been. In other words, the present is path dependent or more simply, history matters. We will consider this effect in more detail in Chapter 18 because it has implications for how we conceptualise an unemployment rate that is consistent with stable inflation.

We will learn that the hysteresis effect describes the interaction between the actual and equilibrium unemployment rates. The implication of hysteresis is that the unemployment rate associated with stable prices at any point in time should not be conceived of as a rigid constraint on expansionary macro policy. The equilibrium rate itself can be reduced by policies that reduce the actual unemployment rate.

For the discussion in this chapter we will confine ourselves to the way the economic cycle impacts on hiring in the labour market.

A recession causes unemployment to rise, and if it is prolonged, short-term joblessness becomes entrenched long-term unemployment, as we noted in the previous section. Thus we would observe a rising average duration of unemployment as the number of long-term unemployed workers rises.

However, the unemployment rate behaves asymmetrically with respect to the economic cycle, which means that it jumps up quickly but takes a long time to fall again.

There is robust evidence to conclude that a worker's chance of finding a job diminishes the longer their spell of unemployment. When there is a deficiency of aggregate demand (and hence lots of unemployed workers seeking jobs), employers use a range of screening devices when hiring. These screening mechanisms effectively 'shuffle' the unemployed queue, with the least desired workers relegated to the back of the queue.

Among other things, firms increase hiring standards (for example, demand higher qualifications than are necessary) and may engage in petty prejudice (screen by age, ethnicity, and so on). A common strategy is to engage in **statistical discrimination**, whereby the firms will conclude that because, on average, a particular demographic cohort shows a particular feature, higher absentee rates (for example), every person from that group must therefore share that negative characteristic. Accordingly, they will screen job applicants using that information even if the average has no relevance to any particular person in that cohort. Personal characteristics such as gender, age, race and other forms of discrimination are thus used to shuffle the disadvantaged workers to the back of the queue.

In this context, the concept of hysteresis relates to how the labour market adjusts over the economic cycle. In a recession, many firms disappear altogether, particularly those which were using very dated capital equipment that was less productive and hence subject to higher unit costs than the best practice technology.

The skills associated with using that equipment become obsolete as it is scrapped, which is more likely to disadvantage the longer-term unemployed. This phenomenon is referred to as skill atrophy. Skill atrophy extends beyond the specific skills needed to operate a piece of equipment or participate in a firm-specific process.

Long-term unemployment also erodes more general skills because the psychological damage of unemployment impacts on a worker's confidence and bearing. A lot of information about the labour market is gleaned informally via social networks and there is strong evidence pointing to the fact that as the duration of unemployment increases, the breadth and quality of an unemployed worker's social network declines.

New entrants to the labour force enter the unemployment pool because of a lack of jobs. They are then denied relevant skills (and the socialisation associated with stable work patterns). Further, because training opportunities are often provided with entry-level jobs, it follows that the (average) skill of the labour force declines when vacancies fall. Thus, both groups of workers, those who have lost their jobs and the new entrants, need to find jobs to update and/or acquire relevant skills. Skill (experience) upgrading also occurs through mobility between jobs, which is restricted during a downturn.

Therefore, workers who have endured shorter spells of unemployment, all else being equal, will tend to be closer to the front of the unemployment queue. Firms assume that those who are enduring long-term unemployment are likely to be less skilled than those who have just lost their jobs and with so many workers to choose from, firms are reluctant to offer training. There is also a self-fulfilling prophecy effect on employer perception – 'This candidate has been unemployed for 18 months, therefore there must be something wrong with them.'

However, just as the downturn generates these skill losses, and imposes longer durations of unemployment on certain groups, a growing economy will start to provide training opportunities and the queue of unemployed diminishes. This is one of the reasons that economists believe it is important for the government to stimulate economic growth when a recession is looming to ensure that skill acquisition can occur more easily.

As demand picks up and the pool of unemployed workers shrinks, employers need to be much less picky. 'Tight full' employment has been defined as a situation in which there are more vacancies than there are unemployed workers seeking jobs. In such conditions, even the most disadvantaged workers can obtain work. Groups that had unfairly faced racial, ethnic, or gender bias and who were at the back of the queue in times of low economic activity, are now able to transition to employment and receive requisite training opportunities. Maintaining tight full employment helps to reduce the likelihood that employers will indulge these irrational biases.

Conclusion

Chapter 5 examined what is often called the labour market or the potential labour force available to an economy and introduced the Labour Force Framework, which classifies labour resources into a number of categories: employed, unemployed, underemployed, and not in the labour force. The chapter discussed how the labour force participation rate, the employment rate, and the unemployment rate are calculated, and demonstrated that any attempt to obtain an accurate measure of a nation's unemployed labour resources is fraught with difficult choices.

Many of those who are officially classified as employed are working for far fewer hours than they would like, and many are working in jobs that do not allow them to fully use their skills and education. Many of those who are officially classified as not in the labour force really would like a job, but have given up looking for one. The chapter examined flows of workers between categories, showing that there is a great amount of flux in labour markets – individuals move among categories so that many of those who lose jobs are not counted among the unemployed, and many of those who gain jobs were not counted as previously unemployed. The use of measures of labour market state transition probabilities helps to demonstrate how dynamic the labour market can be.

Categories of unemployment were examined: frictional, seasonal, structural, and cyclical. The concept of an output gap, which measures the percentage deviation of real GDP from the potential production levels, was introduced. As the output gap rises (that is, as output falls ever further below potential), the unemployment rate rises and labour force participation rates fall.

The chapter also introduced the related concepts of duration of unemployment and hysteresis. When there is a lot of slack in labour markets (the number seeking employment far exceeds the number of job openings), the average time spent unemployed tends to rise. A longer duration of unemployment tends to reduce the likelihood that one will be able to obtain a job – that is, the problem of hysteresis arises: more and more of the unemployed workers come to be seen by potential employers as unemployable.

References

ABS (Australian Bureau of Statistics) (2007) *Labour Statistics: Concepts, Sources and Methods*. Available at: www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/973128B0B872F469CA2572C100244952, accessed 18 January 2017.

ILO (International Labour Organization) (1985) *Labour Statistics Convention (No. 160)*. Available at: http://www.ilo.org/dyn/normlex/en/f?p=1000:12100:0::NO::P12100_INSTRUMENT_ID:312305, accessed 18 January 2017.

ILO (International Labour Organization) (1988) *Current International Recommendations on Labour Statistics*, 1988 edition, ILO, Geneva.

ILO (International Labour Organization) (n.d.) *International Conference of Labour Statisticians*. Available at: https://www.ilo.org/global/statistics-and-databases/meetings-and-events/international-conference-of-labour-statisticians/lang--en/index.htm, accessed 22 August 2018.

US Bureau of Labor Statistics (2016) "Labor Force Status Flows by Sex, January 2016", *Labor Force Statistics from the Current Population Survey*. Available at: http://www.bls.gov/web/empsit/cps_flows_current.htm, accessed 18 January 2017.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

6 SECTORAL ACCOUNTING AND THE FLOW OF FUNDS

Chapter Outline

Learning Objectives

- Develop an understanding of the sectoral balances identity: for every surplus there must be an equal deficit.
- Understand the relationship between sectoral balances and changes to net financial assets.
- Recognise the distinction between vertical and horizontal transactions in their impact on net financial assets.
- Interpret a balance sheet (stocks) and period-to-period changes of its items (flows).

6.1 Introduction

In Chapter 4 we saw that the national accounts divided the national economy into different expenditure categories: consumption by persons/households; investment by private business firms (as well as residential investment in housing); spending by the government; exports to and imports from the foreign sector.

The most basic macroeconomics rule is that one person's spending is another person's income. Another way of stating this rule is that the use of income by one person (that is, spending) will become the source of income for another person or persons.

In this chapter, we extend our understanding of the national accounts, which record these different flows of expenditure and income. The **sectoral balances perspective** of the national accounts brings the uses and sources of national income together. We show that when appropriately defined, the sectoral balances must sum to zero. We expand our discussion of stocks and flows and then introduce the flow of funds by reference to the sectoral balances.

There are many useful insights that can be gained from an understanding of a nation's sectoral balances. The sectoral balances approach helps us to understand the relations between the spending and income balances of the households, firms, government, and foreign sectors of the economy. For example, it illustrates that it is impossible for all sectors to run surpluses (that is, to 'save overall' or spend less than their income) simultaneously. For one sector to run a surplus, we need at least one other to run a deficit (spend more than their income). You will learn that for any of those nations which run external deficits against the rest of the world, in order for its private domestic sector (that is, households plus firms) to run surpluses (that is, to spend less than income and save overall), it is necessary for the government to run fiscal deficits (that is, spend more than it taxes, net of transfers).

6.2 The Sectoral Balances View of the National Accounts

Introduction

The Australian Bureau of Statistics publication *Australian System of National Accounts: Concepts, Sources and Methods* (ABS, 2014) is an excellent resource for understanding the background concepts that are used to derive the sectoral balances framework. The discussion therein is generally applicable to all countries.

From the national accounts sectoral balances framework, economists derived what is called the basic **income-expenditure model in** macroeconomics to explain the theory of income determination that forms the core of the so-called **Keynesian** approach (see Chapter 15).

The income-expenditure model is a combination of accounting identities drawn from the national accounting framework and behavioural theories about how flows of expenditure by households, firms, governments, and foreigners combine to generate sales, which in turn motivate output and income generation.

Remember, that an expenditure **flow** is measured as a certain quantity of dollars that is spent per unit of time. Conversely, a **stock** is measured at a point in time and is the net sum of prior relevant flows.

The accounting aspects that underpin the income-expenditure model draw on several different ways that we can think about the national accounts.

First, from the perspective of the **sources** of national income, we can write out the sources of total spending that flow into the economy over a given period, using the following shorthand:

$$(6.1) \qquad Y \equiv C + I + G + (X - M)$$

Total national income (GDP, represented by Y) is the sum of total final consumption expenditure (C), total private investment (I), total government expenditure (G) and net exports ($X - M$). Note the use of the mathematical symbol $\equiv$, which denotes an **identity** that is true by definition. You should also note that we have seen this identity previously as equation 4.1. You might refresh your memory as to that discussion.

At this stage, we simply take these flows of expenditure as given and understand them to be parts of the national accounts of a nation.

When these components of spending are summed, they equal **aggregate demand for goods and services** in a particular period. Aggregate demand, in turn, generates a response by producers (private and public) in the form of production, which in turn generates flows of income to suppliers of inputs into production (wages, profits). The sum of those flows equals the national income.

As we noted in Chapter 4, the trade account is only one aspect of the financial flows between the domestic economy and the external sector. We must include net external income flows (FNI), which arise from the dividend and income flows that accrue to investments that residents make abroad minus the dividend and interest flows that residents must pay foreign investors who have financial interests within the nation.

Adding in the net external income flows (FNI) to Equation (6.1) for GDP we get the familiar definition of gross national product (GNP):

$$(6.2) \qquad GNP \equiv C + I + G + (X - M) + FNI$$

At this stage, we could make the analysis quite complicated by considering retained earnings in corporations and the like, but here we assume that all income generated by firms and corporations ultimately is received by households, that is, there are no earnings retained by firms.

To obtain the sectoral balances form of the identity, we subtract total taxes net of transfers (T) from both sides of Equation (6.2):

$$(6.3) \qquad GNP - T \equiv C + I + G + (X - M) + FNI - T$$

Now we can collect the terms by arranging them according to the three sectoral balances:

$$(6.4) \qquad (GNP - C - T) - I \equiv (G - T) + (X - M + FNI)$$

The terms in Equation (6.4) are relatively easy to understand now. The term $(GNP - C - T)$ represents total income less the amount consumed by households less the amount paid by households to government in taxes net of transfers. Thus, it represents household saving.

The left-hand side of Equation (6.4), $(GNP - C - T) - I$, thus is the **overall net saving** of the **private domestic sector**, which is distinct from total household saving (S) denoted by the term $(GNP - C - T)$.

In other words, the left-hand side of Equation (6.4) is the **private domestic financial balance** $(S - I)$. If it is positive, then the sector is spending less than its total income (so the sector is adding to its stock of net financial assets), and if it is negative, the sector is spending more than its total income and running down its stock of net financial assets. More generally we define assets as items owned by households and government and non-government organisations which have value. These include financial assets (such as holdings of the money of account, bank deposits and shares) and real assets (such as capital equipment, land and property). On the other hand, liabilities can be held by the same entities and represent financial obligations which need to be settled over time, by some form of payment, which may take a financial form through the transfer of bank deposits or shares, or a real form, namely goods and services.

Note that by rearranging Equation (6.4) we get another version of the sectoral balances equation:

$$(6.5) \qquad (S - I) + (T - G) + (-CAB) \equiv 0$$

The term $(T - G)$ is the **government financial balance** or primary fiscal balance and is in deficit if government spending (G) is greater than government tax revenue (T), and in surplus if the balance is positive.

Finally, the other left-hand side bracketed term $-(X - M + FNI)$ is the negative of the **external financial balance**, commonly known as the current account balance (CAB). It is in surplus if negative and deficit if positive. It is the balance between the spending/income flows of foreigners in the nation and the spending/income flows by residents that go to foreign nations. More simply if the final bracketed term on the left-hand side is positive then there is a current account deficit (CAD).

From Equation (6.5), we can say that:

> **The private domestic financial balance plus the government financial balance plus the current account deficit equals zero.**

This is an accounting statement.

For example, let us assume that the external or foreign balance equals zero. Let us further assume that the private domestic sector's income is $100 billion while its spending is equal to $90 billion, which delivers an overall surplus of $10 billion over the year. Then, from the identity, Equation (6.5), the government sector's fiscal deficit for the year is equal to $10 billion. We know that the private domestic sector will accumulate $10 billion of net financial wealth during the year, consisting of $10 billion of domestic government sector liabilities (given that the external balance is zero).

As another example, assume that there is a current account deficit of $20 billion, so that the spending/income flows from foreigners to the nation is less than the spending/income flows from residents that go to foreign nations. At the same time, assume the government sector spends less than its income, running a fiscal surplus of $10 billion. From our accounting identity, we know that over the same period the private domestic sector must have run an overall deficit equal to $30 billion ($20 billion plus $10 billion). At the same time, its net financial wealth will have fallen by $30 billion because it sold assets and/or issued debt. Meanwhile, the government sector will have increased its net financial wealth by $10 billion (reducing its outstanding debt or increasing its claims on the other sectors), and the foreign sector will have increased its net financial position by $20 billion (also reducing its outstanding debt or increasing its claims on the nation's residents or government).

It is apparent then that, as noted previously, for one sector to run a surplus, at least one other sector must run a deficit. In terms of stock variables, in order for one sector to accumulate net financial wealth, at least one other sector must be in deficit. **It is impossible for all sectors to accumulate net financial wealth by running surpluses.**

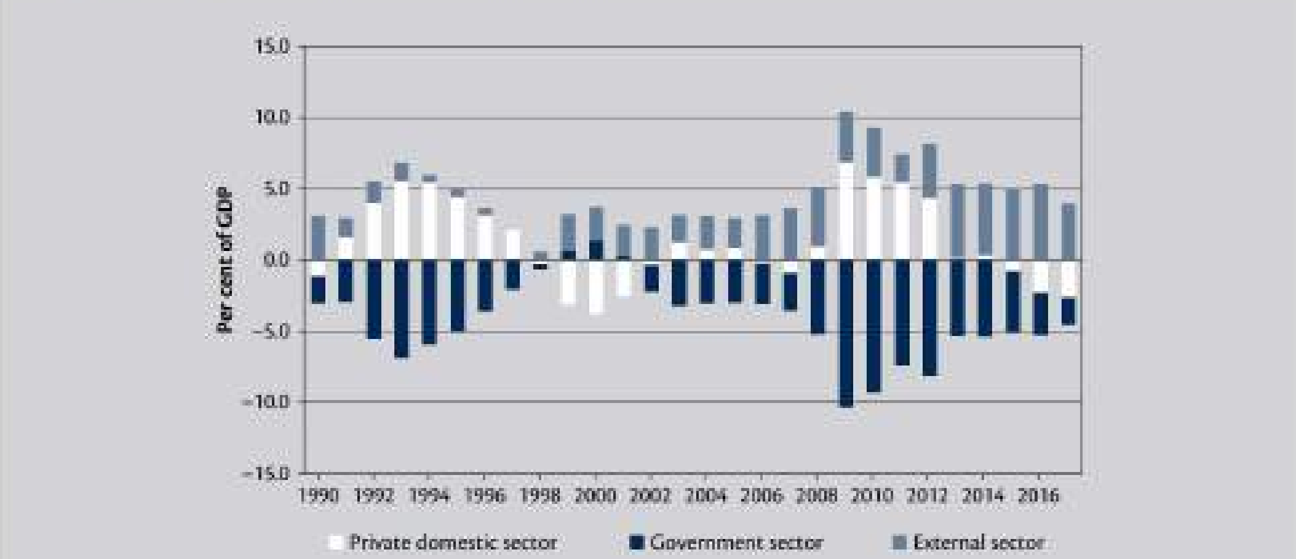
How can we use the sectoral balances framework?

The UK sectoral balances shown in Figure 6.1 replicate Equation (6.5), except that the balances are expressed as percentage shares of GDP. Note that the balances sum to zero.

At this stage three observations are appropriate:

1. Despite the contemporary rhetoric about the desirability of 'getting back to' running an annual fiscal surplus, the UK has rarely done so. Indeed only seven fiscal surpluses have been achieved since 1960, each relatively small and short lived. This is common for other developed nations, too.
2. Like a number of other developed economies, including the US and Australian, current account surpluses have also been relatively rare.
3. Private sector balances have typically been in surplus. The limited occurrences of private sector deficits have been often accompanied by fiscal surpluses. The three annual fiscal surpluses between 1998 and 2000 shown in Figure 6.1 were accompanied by current account deficits and relatively large private sector deficits (7.3 per cent of GDP in 2000). The 2001 economic slowdown followed. In most advanced economies, sharp, severe economic downturns typically follow a period when fiscal surpluses are accompanied by large private sector deficits.

Figure 6.1 UK sectoral balances, 1990 to 2017



Source: Data from OECD (2015).

Note: Imports include net income flows in this graph.

A graphical framework for understanding the sectoral balances

From Equation (6.5) we learned that the sum of the sectoral balances is zero as a matter of accounting. We can construct a diagram defining four quadrants. Figure 6.2 depicts the government fiscal balance on the vertical axis and the external balance on the horizontal axis.

Thus, all points above zero on the vertical axis represent a government fiscal surplus ($T > G$) and all points on the vertical axis below the origin denote government fiscal deficits ($G > T$).

Similarly, all points to the right of the origin on the horizontal axis denote external surpluses ($X + FNI > M$) and all points to the left of the origin on the horizontal axis represent external deficits ($X + FNI < M$). While we shall refer to surpluses and deficits with respect to the sectoral balances, these balances should be understood as being expressed as shares of GDP.

Clearly, the origin of both axes denotes a position where all balances are equal to zero. From Equation (6.5), we also know that when the private domestic balance is zero ($S = I$), then the government fiscal deficit(surplus) will equal the external deficit(surplus). From Figure 6.2, the diagonal 45-degree line thus shows all combinations of government fiscal balances and external balances where the private domestic balance is zero ($S = I$). We will refer to this as the **SI line**.

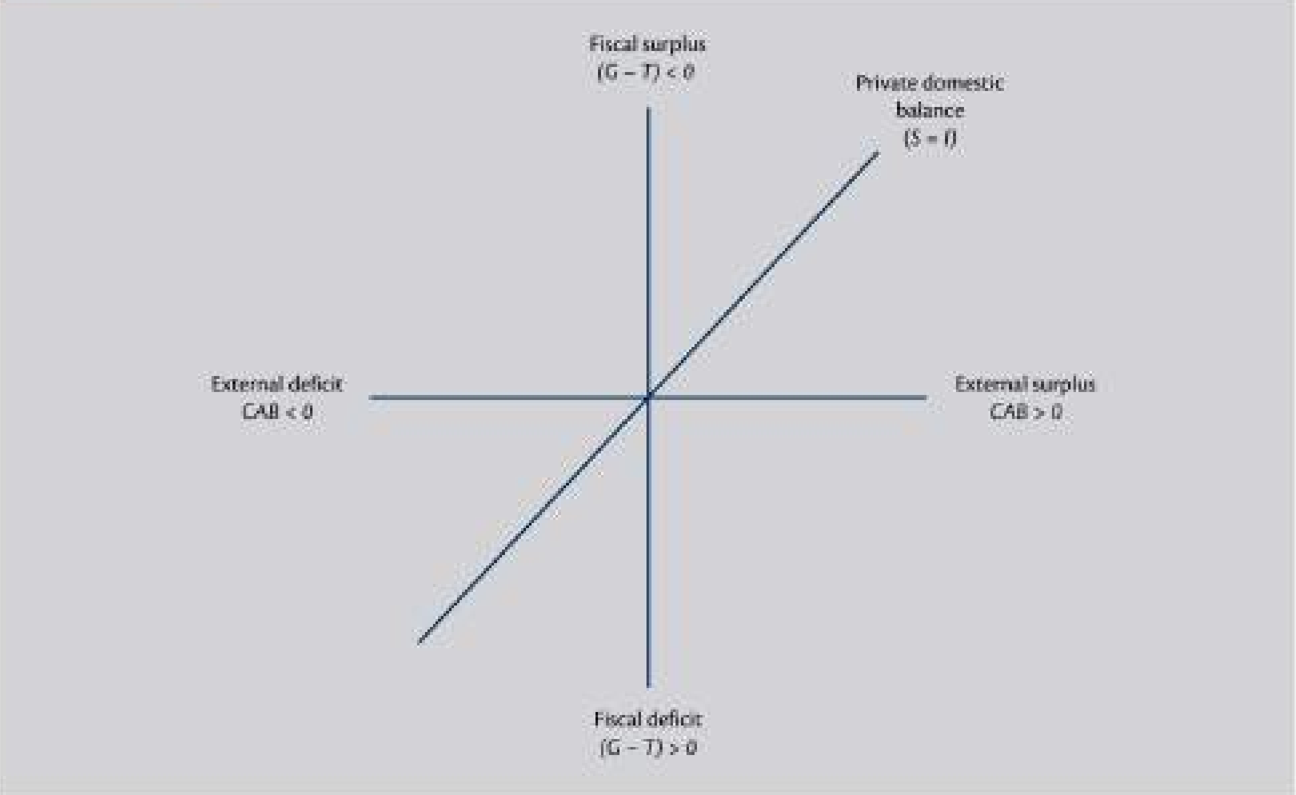
We can use that knowledge to determine the segments of the diagram where the private domestic balance is in surplus ($S > I$) and in deficit ($S < I$). To make it easier, we will use Equation (6.4) re-written in this way:

$$(6.6) \qquad (S - I) = (G - T) + (X - M + FNI) = (G - T) + CAB$$

where $(GNP - C - T) - S$ and $CAB = (X - M + FNI)$. This isolates the private domestic balance on the left-hand side.

We need to identify combinations of the fiscal and external sector balances which yield a private domestic surplus (deficit). At points A and C in Figure 6.3, there is a private domestic balance. Point B corresponds to a

Figure 6.2 A graphical sectoral balances framework



fiscal deficit ($G > T$) and an external surplus ($CAB > 0$). Thus the private sector must be engaging in positive net saving ($S > I$). Then between points B and A, and also B and C, net saving by the private sector is falling until private domestic balance is achieved at points A and C respectively.

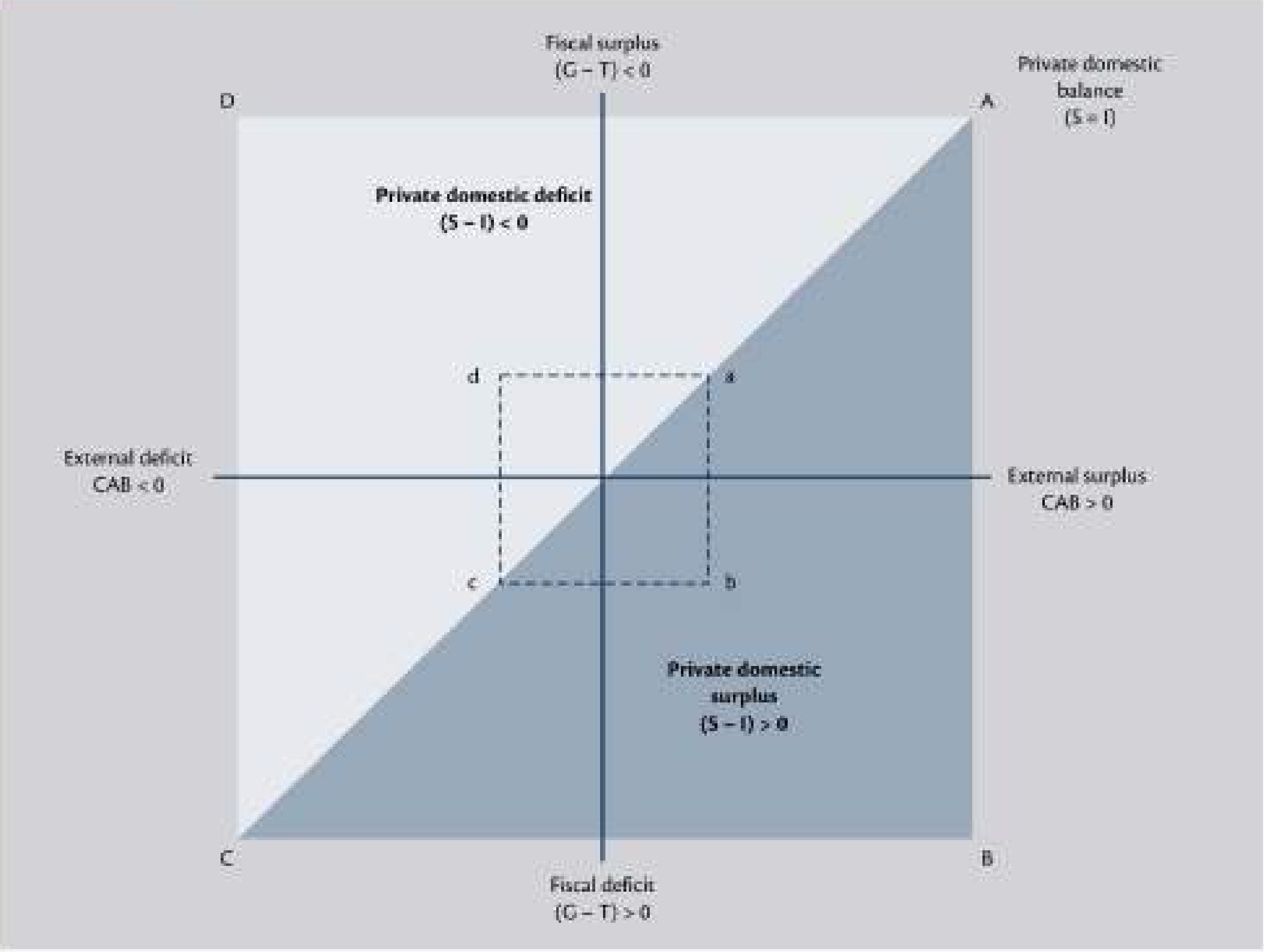
Similarly, it can be readily shown that at D, the private domestic sector is net spending ($S < I$). Between points D and C and D and A, net spending by the private sector declines until private domestic balance is achieved at points A and C respectively.

We can generalise this knowledge and conclude that all points above the 45-degree line on each side of the vertical axis correspond to private domestic sector deficits and all points below the 45-degree line on each side of the vertical axis correspond to private domestic sector surpluses. Consider point b for example, which corresponds to a private domestic surplus, whereas points a and c correspond to a private domestic balance.

The graphical framework thus allows us to examine the implications of different policy options. For a sovereign, currency-issuing government, all combinations of the sectoral balances represented by the points in the four quadrants are permissible. With private sector spending and saving decisions combining with the flows of income arising from trade with the external sector driving national income, the government sector can allow its balance to adjust to whatever magnitude is required to maintain full employment and price stability.

For example, if the external account is in deficit and the private domestic sector is saving overall, then the drain on aggregate demand would require the government to run a deficit of sufficient size to ensure that total spending is sufficient to absorb the real productive capacity available in the economy.

Figure 6.3 Private domestic surpluses and deficits



Alternatively, the external account might be in surplus, which would add to aggregate demand, while the private domestic sector might be spending more than it is earning, that is, in deficit overall. In these situations the government would have to ensure it ran a surplus of sufficient size to ensure that the economy did not overheat and exhaust its productive capacity. The strong economy would be associated with robust tax revenue growth, which would help the government achieve its surplus. Discretionary adjustments in spending and taxation rates might also be required.

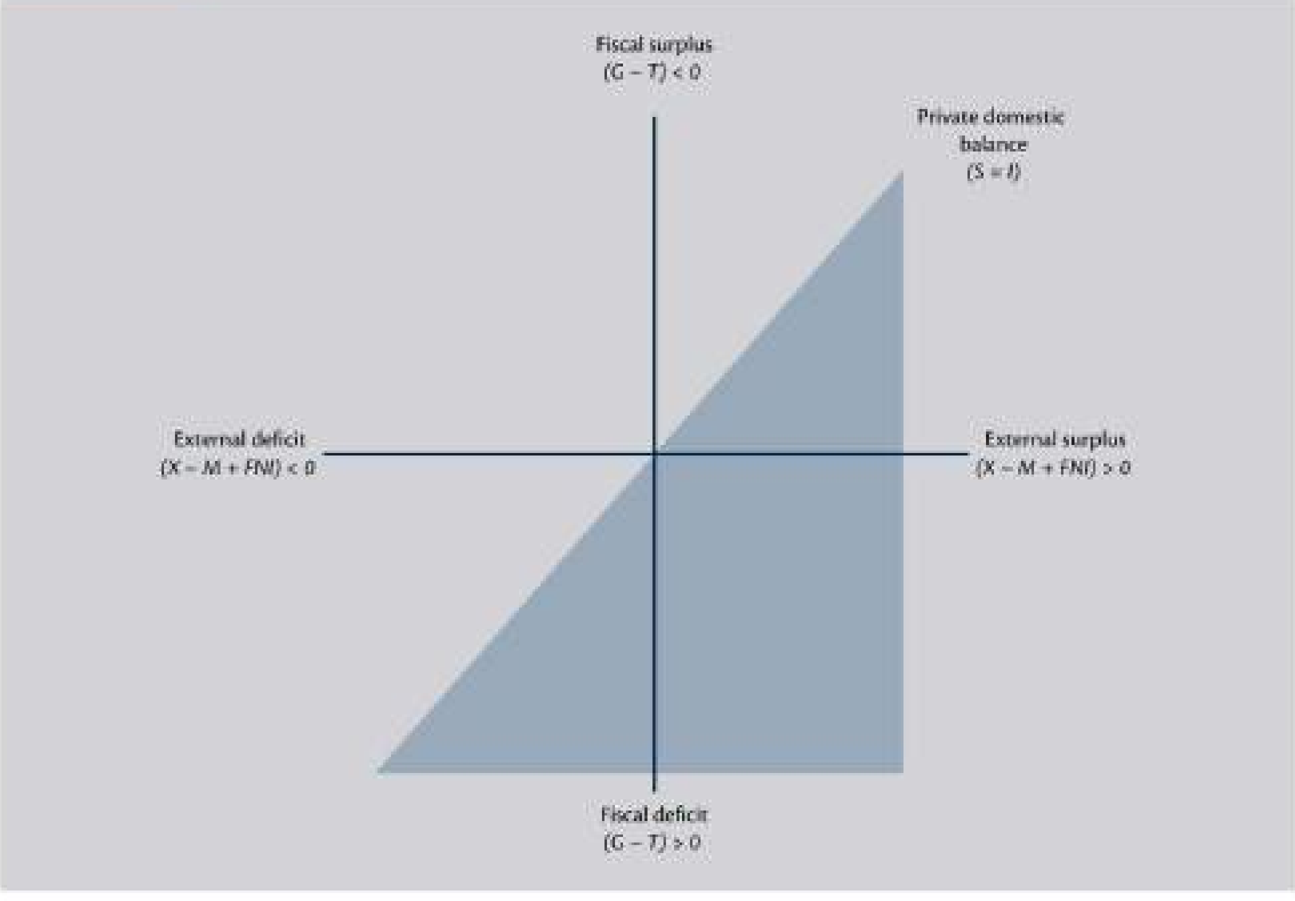
But while these combinations of sectoral balances are permissible, we know that the private domestic sector cannot sustain deficits permanently. This is because the flows of spending which deliver deficits must be funded. As we learned earlier when we considered stocks and flows (see also the discussion in Section 6.4), private domestic deficits ultimately manifest in an increasing stock of debt being held on the private domestic sector's balance sheet.

This process of debt accumulation is limited because at some point the susceptibility of the balance sheet to cyclical movements (for example, rising unemployment) increases and the risk of default rises. In some historical instances, this process has collapsed after serious debt defaults occurred (for example, in the early months of the Global Financial Crisis in 2007–8 (GFC)). At other times, the private sector starts to reduce the precariousness of its balance sheet by reducing spending and increasing saving to bring the debt it is carrying down to a more sustainable level. This will slow economic growth unless it is matched by increased spending by the government or foreign sector.

In the long term, the only sustainable position is for the private domestic sector to be in surplus. An economy can absorb deviations from that position but only for short periods.

Figure 6.4 shows what we might define as the sustainable space available to governments that issue their own currency. Note that this excludes permanent private sector deficits, which are unsustainable.

Figure 6.4 Sustainable space for sovereign governments



Now imagine that the government is forced to operate under a fiscal rule that bans fiscal deficits greater than three per cent of GDP (as shown in Figure 6.5 by the black line). The formation of the European Economic and Monetary Union (Eurozone) introduced just such a fiscal rule under its Stability and Growth Pact. The aim was to restrict the capacity of each member state to run government fiscal deficits.

What does such a fiscal rule mean for both permissible and sustainable spaces available to a macroeconomic policymaker?

Clearly any point above the three per cent of GDP fiscal deficit line in Figure 6.5 is permissible. However, using the same logic as before, the sustainable space requires that the private domestic sector be in surplus overall, even though short-term deviations from this can occur from time to time.

Figure 6.5 shows the sustainable space for such an economy (the combination of grey and blue areas). The blue shaded area shows the sustainable space available to policymakers in nations that run external surpluses. The grey shaded area shows the sustainable space available to policymakers in nations that run external deficits.

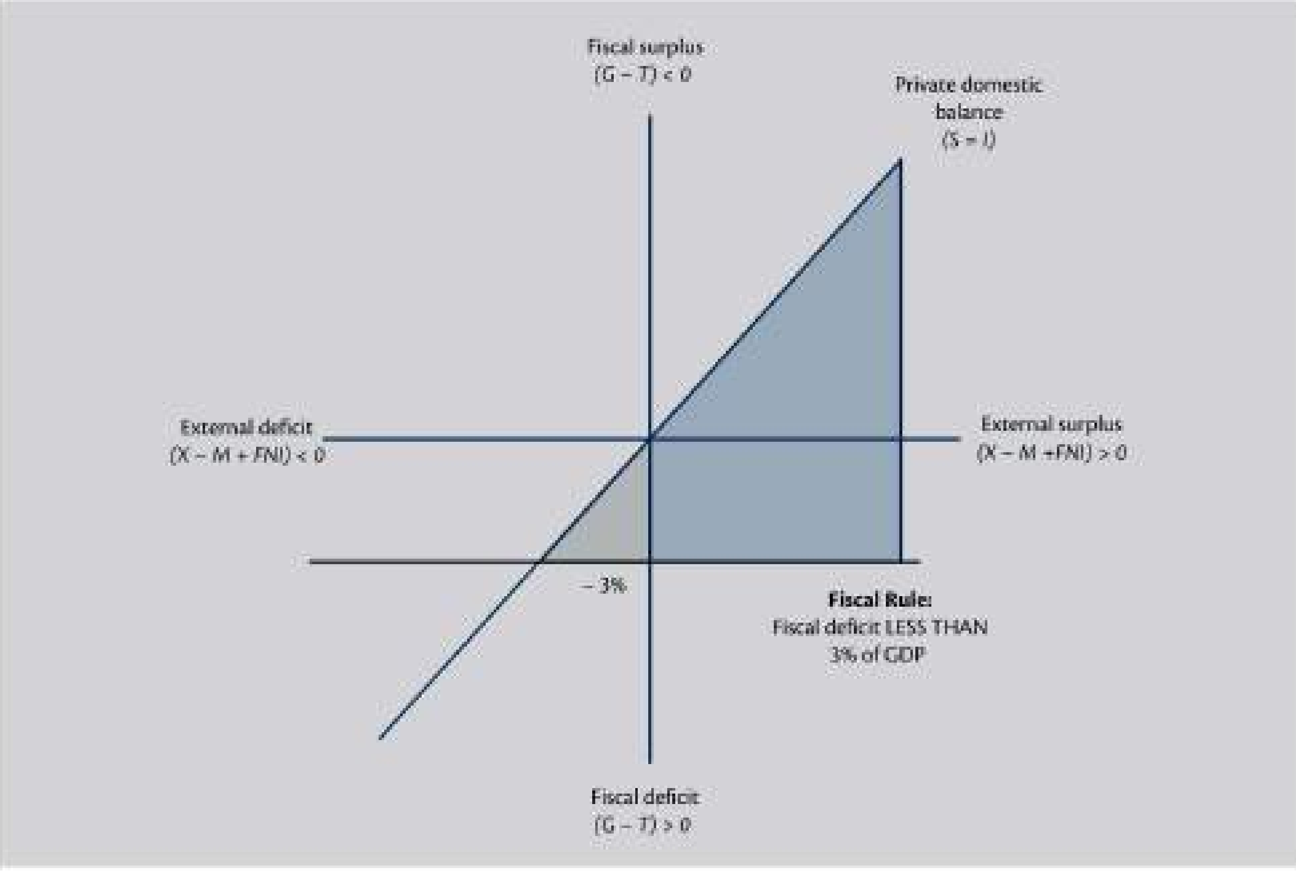
Thus, the policy space that governments have to operate within when fiscal rules are imposed is very limited relative to the options available to a sovereign, currency-issuing government, operating without any direct quantitative restrictions on the deficits they can run.

Why is this important? An unconstrained government can always utilise the available space to ensure aggregate demand is sufficient to maintain full employment and price stability.

By definition, not every nation can run an external surplus because an external surplus in one nation has to be matched by external deficit(s) in other nations. While the external surplus nations have more policy flexibility when operating under a fiscal rule of the type shown in Figure 6.5, the fact remains that the allowable fiscal deficits may be insufficient to maintain the aggregate demand necessary to sustain full employment.

The policy inflexibility facing nations which run external deficits and simultaneously have to operate under fiscal rules becomes even more restrictive, as shown in Figure 6.5 by the small grey triangle. When such an economy

Figure 6.5 Sustainable space for governments constrained by fiscal rules



experiences a negative economic shock significant enough to drive the private sector to reduce its spending and target a sectoral surplus, the extent to which the fiscal deficit can be used to absorb the loss of overall aggregate demand is very limited.

It is highly likely that such an economy will experience enduring recessions as a result of the artificial fiscal rules (restrictions) that are placed on its government. Note that such a situation befell Greece and some other Mediterranean European nations after the GFC impacted.

Note that Figures 6.2–6.5 show the feasible combinations of sectoral balances under the constrained and unconstrained scenarios, but do not show which combinations are associated with full employment.

REMINDER BOX

The sustainable goal for a government should be to maintain full employment and price stability and allow its fiscal balance to adjust accordingly to ensure aggregate demand is consistent with those goals. A sovereign, currency-issuing government can always meet those goals if it chooses.

However, the imposition of fiscal rules restricts the government from achieving these goals and makes a pre-determined fiscal outcome the target of macroeconomic policy rather than the more significant macroeconomic objectives of full employment and price stability.

The lesson here is that the government should never specifically target any particular fiscal outcome, but rather, should target employment growth and price level stability.

6.3 Revisiting Stocks and Flows

Flows

In this section we re-examine the concepts of stock and flow variables, which were briefly outlined in Chapter 1. This will enable us to clearly set out the relationships between deficit spending and saving, and between financial deficits and debt. This section clarifies these fundamental accounting relationships.

Flow variables are measured over time. The simplest example is personal income, which can be stated as $10 per hour, or $400 per week, or $20,000 per year. The important point is that without a clear statement of the time component, any statement about a flow is incomplete and somewhat meaningless. If one says one's income is $100, we need to know whether that is per hour, per day, per week, or per year to make sense of it. It is also useful to work with the growth of flow variables, often calculated as annual growth rates. For example, your employer might offer a labour contract that provides for annual cost of living increases equal to four per cent per year. In the first year, you would receive $20,000, while in the second you would receive a wage income of $20,800 ($20,000 plus 4% of $20,000, which is equal to $800).

What flows? When we speak of the flow of a river, it is obviously water that is flowing, measured in terms of thousands of cubic metres per second. However, it is not so clear what is flowing when we discuss flows of income and expenditure. For example, what 'flows' to provide a wage income equal to $20,000 per year? The simple answer is 'dollars'. You work for your employer for eight hours a day, five days a week, and after two weeks you receive an electronic transfer for the sum of $800 (ignoring here for simplicity any possible deductions for taxes and benefits). Even on payday, it is difficult to think of the electronic transfer as the 'dollars' that were flowing while you were working. As we will see in Chapter 9, the payment (say, in the form of an electronic transfer) is just an IOU issued by your employer's bank that is denominated in your nation's money of account; dollars in our example.

In fact, we can conceive your work for hourly wages as an implicit accumulation of the IOUs of your employer. Over the course of the two weeks during which you worked, you earned a flow of wages equal to ten dollars for each hour worked, received in the form of an implicit promise from your employer to pay you in dollars at the end of the two-week period. Indeed, in the event of a dispute, the court system would recognise the legal obligation of your employer to pay dollars to you for hours worked. In this sense, we can think of each hour worked leading to your accumulation of the IOUs of your employer as being denominated in dollars. On payday, your employer extinguishes the IOUs by delivering to you a funds transfer for the total obligations accumulated over the two-week period. Two important conclusions follow from this example.

Flows are measured in terms of money. The money of account is how we measure flows of income or spending. The associated flow of currency can take the physical form of notes and coins, but equally can be an electronic deposit, say in a private bank account. Thus, in contrast to a flow of water, the flows of spending or income do not always take a physical form.

As we will explore later, metal coins and paper currency are nothing more than government IOUs denominated in the money of account. While government currency differs in some respects from the implicit IOUs you accumulate against your employer, all share the common characteristic that all are IOUs, denominated in dollars.

We also need to differentiate between flows of income and spending denominated in the money of account from the associated flows of (labour) services and consumer goods and services. In principle, consumer goods and services are used up to satisfy the needs and desires of households. However, consumption purchases made this week could include goods that will be used for many months or even years. Economists typically record consumption at the time the purchase is made and at the dollar value of the purchase even while recognising that goods and services purchased might provide a stream of 'satisfaction' over a longer period of time.

Stocks

Flows accumulate as stocks. The flow of water in a stream can be accumulated in a reservoir behind a dam, or in the cup we dip into the stream. The stock of water is then the number of cubic metres in the reservoir, or the half litre in the cup. Unlike a flow, a stock can be measured without reference to a time period because it exists at a point of time. We can measure the stock of water in a lake at noon on the last day of the summer as 1.5 billion cubic metres, and at noon on the last day of the following winter as 2.0 billion cubic metres. Because the stock has increased, we can surmise that the inflow of water during the passing of six months has been greater than the outflow of water over that period, by an amount of 0.5 billion cubic metres.

Let us continue to assume that you receive a biweekly wage payment in the form of an electronic transfer for $800, 25 times a year, for a total annual income of $20,000 (note the slight inaccuracy with respect to a standard calendar year – we want to keep the numbers easy). On each payday, the electronic transfer appears in your bank account, increasing your deposits by $800. Your bank deposit represents a portion of your wealth, held in the form of a financial asset, which is a claim on your bank. (We examine the properties of financial assets in Chapter 10.) Because wealth is measured at a point in time, it is a stock variable. In addition to your bank account, you might also hold other forms of financial wealth (stocks and bonds, currency in your pocket, other types of bank deposits) as well as real wealth (a car, real estate, a business firm, art and jewels). Again, these are stock variables whose value is measured in terms of the money of account at a point in time.

Once you have received the $800 transfer, you begin to draw down your bank account to finance your purchases. Let us assume that your annual consumption will be $18,000 for the year, comprising purchases of consumer goods (food, fuel for your automobile, clothing) and consumer services (entertainment, medical care, legal services). Hence, between wage payments, you spend a total of $720 for consumption, drawing down your bank account by that amount to finance these purchases.

Over the year, your flow of wage income has been equal to $20,000 and you have spent $18,000 of that on consumption. Your flow of saving over the year is also equal to $2,000, because saving is defined as the residual dollar value of income that has not been spent over the period.

This will accumulate as an addition to your stock of wealth. If you allow the funds to accumulate in your bank account (which we will initially assume does not earn interest), the annual addition to your financial wealth will be $2,000. Alternatively, you could purchase interest earning bonds, which is another form of financial wealth. In this case, however, you will also have a flow of interest earnings in addition to your labour income. The flow of interest income (let us say it amounts to $200 over the course of the year) will also add to your stock of financial wealth so that the total addition to your stock of financial wealth will be $2,200 over the year.

However, there are many other possible uses of your saving flow. You might decide to buy stocks or other kinds of financial assets. Or, you might purchase real assets: a classic car, real estate, or equipment for your family's business. Your saving decision can be analysed as a two-step process: first as a decision to withhold a portion of your income flow from spending, and second a decision as to the form in which your wealth will be accumulated. An income flow is first realised as an accumulation of IOUs, normally claims on a bank in the form of a deposit, that in the second step is used to purchase an asset, which might be another financial IOU, or a real asset.

One person's financial asset is another's financial liability. It is a fundamental principle of accounting that for every financial asset there is an equal and offsetting financial liability. The bank deposit account is a household's financial asset, offset by the bank's liability (or IOU). A government or corporate bond is a household asset, but represents a liability of the issuer (either the government or the corporation). The household probably has some liabilities too, such as student loans, a home mortgage, or a car loan. These are held as assets by the creditor, which could be a bank or any of a number of types of financial institutions including pension funds, hedge funds, or insurance companies. A household's net financial wealth is equal to the sum of all its financial assets (equal to its financial wealth) less the sum of its financial liabilities (all of the money-denominated IOUs it issued). If the overall sum is positive, the household has positive net financial wealth.

Examples of stocks include stock of capital, inventories, financial wealth, and net worth.

Inside wealth versus outside wealth

It is often useful to distinguish between types of sectors in the economy. The most basic distinction is between the public sector (including all levels of government) and the domestic private sector (households and firms). Note that we are simplifying by excluding the foreign sector as if the economy were completely closed to trade and capital flows.

If we were to take all of the privately issued financial assets and liabilities, it is a matter of logic that the sum of financial assets must equal the sum of financial liabilities. In other words, net financial wealth would have to be zero if we consider only private sector IOUs. This is sometimes called 'inside wealth' because it is 'inside' the private sector. In order for the private sector as a whole to accumulate net financial wealth, the inflow must be in the form of 'outside wealth', that is, financial claims on another sector. Given our basic division between the public sector and the domestic private sector, the outside financial wealth takes the form of government IOUs. The private sector holds government currency (including coins and paper currency) as well as the full range of government bonds (short-term bills, longer maturity bonds) as net financial assets, which is a portion of its positive net wealth.

Net private financial wealth equals public debt in our closed economy without a foreign sector. Recall from our discussion above that the accumulation of stocks requires flows. The private sector's accumulation of net financial assets over the course of a year is made possible only because its spending is less than its income over that same period. In other words, it has been saving, which enables it to accumulate a stock of wealth in the form of financial assets. In our simple example with only a public sector and a domestic private sector, these net financial assets are government liabilities – government currency and government bonds. These government IOUs, in turn, can be accumulated only when the government spends more than it receives in the form of tax revenue. This is called the fiscal deficit, which is the flow of government spending less the flow of government tax revenue measured in the money of account over a given period (usually a year). This deficit accumulates to a stock of government debt, which will be equal to the private sector's accumulation of financial wealth over the same period.

A complete explanation of the process of government spending and taxing will be provided in Chapter 20. What it is necessary to understand at this point is that in our two-sector example the net financial assets held by the private sector are exactly equal to the net financial liabilities issued by the government. If the government's spending always equals its tax revenue, the private sector's net financial wealth would be zero.

Rest-of-world debts are domestic financial assets. We can broaden our analysis by considering the financial assets and liabilities of the rest of the world. Thus, we now form three sectors in this open economy: a domestic private sector, a domestic public sector, and a 'rest of the world' sector that consists of foreign governments, firms, and households. In this case, it is possible for the domestic private sector to accumulate net financial claims on the rest of the world, even if the domestic public sector's spending over the period exactly equals its tax revenue. The domestic sector's accumulation of net financial assets is equal to the rest of the world's issue of net financial liabilities. Finally, and more realistically, the domestic private sector can accumulate net financial wealth consisting of both domestic government liabilities and rest-of-world liabilities. It is also possible for the domestic private sector to accumulate government debt (adding to its net financial wealth) while also issuing debt to the rest of the world (reducing its net financial wealth).

Non-financial wealth (real assets)

One's financial asset is necessarily offset by another's financial liability. However, real assets represent one's wealth that is not offset by another's liability so at the aggregate level net wealth equals the value of real (non-financial) assets. To be clear, you might have purchased an automobile by going into debt. Your financial liability (your car loan) is offset by the financial asset held by the auto loan company. Since that asset and liability net to zero, what remains is the value of the real asset, the car. In most of the discussion that follows we will be concerned with financial assets and liabilities, but we will keep in the back of our minds that the value of real assets provides net wealth at both the individual level and at the aggregate level. Once we subtract all financial liabilities from total assets (real and financial) we are left with non-financial (real) assets, or aggregate net worth.

6.4 Integrating NIPA, Stocks, Flows and the Flow of Funds Accounts

The sectoral balances framework, which is derived from the national accounts framework, was explored in Section 6.2. It is intrinsically linked to the **flow of funds analysis**. They are different, but related, ways of considering national economic activity.

An early exponent of the flow of funds approach, Lawrence Ritter (1963: 220) wrote that:

> *The flow of funds is a system of social accounting in which (a) the economy is divided into a number of sectors and (b) a 'sources-and-uses-of-funds statement' is constructed for each sector. When all these sector sources-and-uses-of-funds statements are placed side by side, we obtain (c) the flow-of-funds matrix for the economy as a whole.*

Thus, the flow of funds accounts allow us to link a sector's balance sheet (statements about stocks of financial and real net wealth) to income statements (statements about flows) in a consistent fashion. In a monetary economy, flows of expenditures measured in terms of the money of account spent over a period involve transactions between sectors in the economy, which also have logical stock counterparts; that is, flows feed stocks. The flow of funds accounts ensure that all of these transactions are correctly accounted for.

This thinking underpinned the work of the so-called New Cambridge approach that was part of the Cambridge Economic Policy Group at the University of Cambridge in the early 1970s. Key members of this group were Martin Fetherston, Wynne Godley and Francis Cripps, all of whom were of a Keynesian persuasion.

While the sectoral balances approach had been understood much earlier (for example, by Nicholas Kaldor and others), it was popularised by the New Cambridge macroeconomic analysis which put the concept of the **net acquisition of financial assets (*NAFA*)** into the forefront of its Keynesian income-expenditure model.

Like Lawrence Ritter, the Cambridge economists were interested in tracing the flow of funds between the different sectors of the economy, which they divided into the government sector, the private domestic sector and the external sector. These transactions are recorded for a given period, and each sector could record a financial deficit or surplus.

We can rewrite Equation (6.6) as follows:

$$(S - I) = NAFA = (G - T) + CAB \qquad (6.7)$$

$(S - I)$ is the private domestic financial balance or *NAFA* of the private domestic sector. The private domestic sector is in financial surplus(deficit) when its disposable income $(GNP - T)$ exceeds(is less than) its spending on consumption goods and investment goods.

From a stock perspective, *NAFA* can also be measured by the difference between the private domestic sector's stock of net financial assets at time $t - 1$ and the stock at time t, so if t is 2017, $t - 1$ would be 2016.

If $G - T < 0$, then the government sector is spending less than it is taking out of the economy in taxation and undermining the capacity of the other two sectors to accumulate net financial assets by running surpluses and vice versa.

CAB is the external sector financial balance (the current account balance) and comprises the **trade balance** (that is, the difference between export and import revenue on goods and services) and the **net income flows** that accrue to residents as a consequence of interest and dividends received on overseas ownership (offset by similar payments to foreigners).

If the overall external sector balance is in deficit, then the national economy is issuing liabilities abroad or running down its net financial position in other ways, and foreigners are accumulating financial asset claims (and vice versa when the external balance is in surplus).

Noting the stock/flow distinction, Equation (6.7) can be interpreted as meaning that if its right-hand side is positive, then $(G - T) + CAB > 0$, and the government sector fiscal balance plus the current account balance jointly generate national income and additional net financial assets for the private domestic sector. Then $NAFA > 0$, which means that the private domestic sector is running a surplus and acquiring new assets and/or reducing its existing debt obligations.

Conversely, if the government sector fiscal balance plus the current account balance is negative, this would reduce national income and undermine the capacity of the private domestic sector to net save and add to its stock of net financial assets. In this case $NAFA < 0$, so that the private domestic sector is running down its net financial position by borrowing from the other sectors and/or by liquidating some of its stock of accumulated wealth.

Equation (6.7) can also be written as:

$$[(S - I) - CAB] = (G - T) \tag{6.8}$$

where the term on the left-hand side $[(S - I) - CAB]$ is the non-government sector financial balance and is of equal and opposite sign to the government financial balance, $(G - T)$.

This is the familiar MMT conclusion that a government sector deficit (surplus) is equal dollar for dollar to the non-government sector surplus (deficit).

Importantly, **transactions within the private domestic sector do not alter the net financial position of that sector overall**. So, if a bank creates a loan for one of its customers then its assets rise but on the other side, the liabilities of the customer increase by an equal amount, leaving no change in the net position of the sector.

The only way the private domestic sector can increase its net financial assets overall is through transactions with the government or external sector, for example, by acquiring a government bond or buying a foreign government bond (or a foreign corporate bond). These two points are key MMT insights.

Once we understand the interlinked nature of the three sectors, then it is a simple step to realise that if one sector has improved its position by the net acquisition of financial assets, following a financial surplus, at least one other sector must have reduced its net financial assets or run a financial deficit.

The flow of funds framework allows us to understand that the funds which a particular sector receives during a period (from current receipts, borrowing, selling financial assets, and running down cash balances) have to be equal to the total of its current expenditures, capital expenditures, debt repayments, lending, and accumulation of cash balances. This approach clearly allows us to trace the uses and sources of funds for each sector.

It should be emphasised that the flow of funds approach is based on accounting principles rather than being a behavioural (theoretical) framework for understanding why these flows occur. Nor do we gain any insights as to the adjustment processes that govern the change in net financial assets in each sector. That caveat is not to be taken as a criticism of the approach; but merely an observation of its restrictions. It also doesn't reduce the utility and insights that the approach provides. Often economists like to denigrate analyses that manipulate accounting identities as being too low brow. But any approach is valuable if it provides useful ways of thinking.

Causal relationships

From the discussion above, it is clear that a non-government surplus is the same thing as a saving flow which leads to the net accumulation of financial assets. By the same token, a deficit reduces net financial wealth. If the private domestic or external sector runs a deficit, it must either use financial assets that it has accumulated in previous years (when surpluses were run), or it must issue new IOUs to offset its deficits (that is, borrow).

Thus, the sector 'pays for' its deficit spending by selling assets and reducing its bank deposits ('dissaving'), or it borrows (issues debt) to obtain bank deposits. Once it runs out of accumulated assets, it has no choice but to increase its indebtedness every year that it runs a deficit. On the other hand, if the external or private domestic sector runs a surplus then it will be accumulating net financial assets. This will take the form of financial claims on at least one of the other sectors.

As we will discuss later, it is misleading to apply terminology such as 'dissaving' or 'borrowing' to the sovereign government, which issues the currency.

While we have identified an accounting relationship between the sectoral balances, we can also say something about the causal relationships between the flows of income and expenditure and the impact on stocks.

Individual spending is mostly determined by income. For the individual, it is plausible to argue that income determines spending because an individual with no income is certainly going to be severely constrained when deciding to purchase goods and services. However, on reflection it is apparent that even at the individual level, the link between income and spending is loose; one can spend less than one's income, accumulating net financial assets, or one can spend more than one's income by issuing financial liabilities and thereby becoming indebted. Still, at the level of the individual household or firm, the direction of causation runs from income to spending even if the correspondence between the two flows is not perfect.

Deficits create financial wealth

We can also say something about the direction of causation regarding accumulation of financial wealth at the level of the individual. If a household or firm decides to spend more than its income by running a deficit, it can issue liabilities to finance the additional purchases. Another household or firm will accumulate these liabilities as net financial wealth. Alternatively, they might allow the government to run a fiscal surplus. Of course, for this net financial wealth accumulation to take place, we must have one household or firm willing to deficit spend, and another household, firm, or government willing to accumulate wealth in the form of the liabilities of that deficit spender-'it takes two to tango'. However, the decision to deficit spend is the initiating cause of the creation of net financial wealth. No matter how much others might want to accumulate financial wealth, they will not be able to do so unless someone is willing to deficit spend. Still, it is true that the household or firm will not be able to deficit spend unless it can sell some of its accumulated assets or find someone willing to hold its liabilities, such as a bank through the creation of a loan.

In the case of a sovereign government, there is a special power – the ability to tax – that guarantees that households and firms will want to accumulate the government's debt. We conclude that while causation is complex, it tends to run from individual deficit spending to accumulation of financial wealth, and from debt to financial wealth. Since the accumulation of a stock of financial wealth results from a surplus, that is, from a flow of saving, we can also conclude that causation tends to run from deficit spending to saving. At the sectoral rather than individual level the same principles apply. Thus, one sector cannot run a deficit if no other sector will run a surplus. Equivalently, we can say that one sector cannot issue debt if no other sector is willing to accumulate the debt instruments.

At the aggregate level, taking the economy as a whole, causation is clearer. A society cannot decide to have more income, but it can decide to spend more. Further, all spending must be received by someone, somewhere, as income. In other words, aggregate spending creates aggregate income. Finally, as discussed above, spending is not necessarily constrained by income because it is possible for households, firms, or government to spend more than income. Indeed, any of the three main sectors can run a deficit with at least one of the others running a surplus. However, it is not possible for spending at the aggregate level to be different from aggregate income since the sum of the sectoral balances must be zero. For all of these reasons, we must reverse causation between spending and income when we turn to the aggregate level: while at the individual level income causes spending, at the aggregate level spending causes income.

In MMT, we differentiate between horizontal and vertical transactions within the economy. Horizontal transactions occur between people and firms within the non-government sector (for example, purchases of goods and services, borrowing from banks). Vertical transactions occur between the government sector and the non-government sector (for example, government spending and taxation).

Horizontal transactions do not add to the stock of net financial assets held by the non-government sector. Much of the debt issued within a sector will be held by others in the same sector. For example, if we look at the finances of the private domestic sector we will find that most business debt is held by domestic firms and households. In the terminology we introduced above, this is 'inside debt' of those firms and households that run budget deficits, which is held as 'inside wealth' by those households and firms that run budget surpluses. Likewise, if households choose to deficit spend, that is, spend more than their flow of annual income, then they may secure bank loans. In this case the net asset position of the private sector is unchanged. These are horizontal transactions. Student loans may be privately financed or financed by government. In both cases there is a horizontal transaction. If the domestic private sector taken as a whole spends more than its income, it must issue 'outside debt' held as 'outside wealth', typically held by the foreign sector. However, the stock of net financial assets held by the non-government sector (private domestic plus foreign) is again unchanged, since these are horizontal transactions.

The initiating cause of the private sector deficit is assumed to be a desire to spend more than income, so the causation mostly goes from deficits to surpluses and from debt to net financial wealth. While we recognise that no sector can run a deficit unless another wants to run a surplus, this is not usually a problem because there is a propensity to net save financial assets.

Vertical transactions add to the stock of net financial assets held by the non-government sector. On the other hand, assume that a fiscal deficit occurs (perhaps as a result of increased government spending), and for simplicity the CAB is zero, then the private sector achieves a net increase in its stock of financial assets. This transaction between the government and private sector is referred to as a vertical transaction, and in this instance, leads to an increase in net financial assets held by the non-government sector. On the other hand, if the government runs a fiscal surplus (by taking net spending out of the economy), with the CAB again being zero, the non-government sector (specifically the private sector) suffers a loss in its net holdings of financial assets.

In this section, we demonstrate how a flow of funds approach to the analysis of monetary transactions highlights both the importance of the distinction between vertical and horizontal transactions and the fundamental accounting nature of the so-called government 'budget' constraint (GBC) identity, which we will refer to as the government fiscal constraint.

6.5 Balance Sheets

Following Ritter, we can present a very simple 'generalised balance sheet', which would apply to any sector, as being depicted in the T account shown in Figure 6.6. By a T account we are referring to a set of financial records that uses double-entry bookkeeping.

Figure 6.6 A stylised sectoral balance sheet

Assets	Liabilities and net worth
Financial assets:	Liabilities
1. Money	
2. Other	
Real assets	Net worth
Σ	Σ

Several points are worth noting. Real assets only appear on the balance sheets of their owners. Financial assets are different to real assets because they represent the indebtedness of other sector(s). This means that they will be matched by a financial liability on at least one other sector's balance sheet.

Financial assets denote monetary amounts owned by that sector, which by the same logic as before means that there will be a matching liability on at least one other balance sheet within the system.

When we consider the monetary system as a whole, we conclude that financial assets and financial liabilities net to zero: that is, the total value of the financial assets equals the total value of outstanding liabilities.

This accounting also tells us that for the overall economy, net worth equals the monetary value of the real assets in the economy.

The balance sheet depicts stocks but we can easily see how it might provide us with information about flows, in the way the national accounts do. A stock is measured at a point in time (say, the end of the year) whereas flows measure monetary transactions over a period (say, a year).

If we examine the difference between a balance sheet compiled at, say, 31 December 2017, and a balance sheet compiled at 31 December 2018, we will be able to represent the information in the balance sheet about assets, liabilities and net worth as flow data.

Consider Figure 6.7 (where the Δ symbol refers to changes over the period concerned). Now the entries in the T account denote uses and sources of funds (that is, flows) over the period of interest. There are two components, one relates to real assets and net worth, and the other to financial assets.

A given sector (for example, household, firm, government) can in the first instance obtain funds by increasing their liabilities through borrowing and incurring debt (ΔL). They can apply those funds to accumulating more financial assets (ΔFA) or building cash balances (ΔM).

If we wanted to complicate matters we could decompose ΔFA, ΔM and ΔL further, by recognising that a given sector can also sell existing financial assets or run down cash balances to obtain new funds. Similarly, it might use funds to reduce liabilities (pay down debts). Thus the entries in Figure 6.7 are to be considered as **net transactions**.

The second source and use of funds for a sector relates to changes in real assets (ΔRA) and the change in net worth (ΔNW) over a given period.

In the national accounts framework (see Chapter 4), we considered the division between the capital account and the current account, where the former related to investment in productive capacity and the latter referred to recurrent spending and income. The capital account measured transactions, which change the real assets held and the net worth of the economy.

Figure 6.7 A uses and sources of funds statement

Uses	Sources
Δ Financial assets (lending)	Δ Liabilities (borrowing)
Δ Money (cash balances)	
Δ Real assets (investment)	Δ Net worth (saving)
Σ	Σ

What do we mean by a change in real assets? In the national accounts, we considered gross capital formation or investment, which is defined as expenditure on productive capital goods (for example, plant and equipment, factories). This is a use of funds by firms in the current period. Depreciation represents the difference between gross and net investment. For now, though, we abstract from that real world complexity of depreciation.

Finally, we consider the change in net worth for a sector in a given period to be the residual after all the uses and sources of funds have been accounted for. From an accounting perspective, net worth is equal to the difference between total assets and total liabilities.

It follows that a change in net worth over the period of interest is equal to the difference between the change in total assets and the change in total liabilities. If total assets increase by more(decrease by less) than total liabilities increase(decrease) then the net worth of the sector has risen.

Another way of thinking about the change in net worth, which is a flow of funds, is to link it to the national accounts concept of saving.

In the national accounts framework, we consider household saving, for example, to be the difference between consumption (a use) and disposable income (a source). This concept generalises (with caution) to the statement that the surplus of a sector is the difference between its current revenue and its current expenditure.

What happens to the flow of surplus funds? If the current flow of income is greater than the current expenditure, then at the end of the period, the sector would have accumulated an increased stock of total net assets, by increasing the actual assets held and/or reducing liabilities owed.

The surplus between current income and current expenditure must be matched dollar for dollar by an increase in the stock of total net assets. We have already discussed total net assets above but in different terms.

We defined the change in net worth over a period as the difference between the change in total assets and the change in total liabilities. That difference is exactly equal to the surplus of current income over current expenditure.

Thus, from an accounting perspective, we can consider saving to be the change in net worth over a period.

Figure 6.7 however only implicitly includes the current account transactions, the flow of current income and expenditure, since we defined the change in net worth (ΔNW) to be the difference between the two current flows.

The simplicity of Figure 6.7 is that it shows that if a sector is running a deficit (that is, it is spending more than it is earning or in the parlance used above, it is investing more than it is saving) then it must obtain the deficit funds from its available sources:

- Increasing borrowing ($\Delta L > 0$)
- Running down cash balances ($\Delta M < 0$)
- Selling existing financial assets ($\Delta FA < 0$)
- Selling existing real assets ($\Delta RA < 0$)

More generally, a given sector (for example, household, firm, government) may be running a deficit, but could choose to reduce their liabilities ($\Delta L < 0$) by running down their stocks of financial assets ($\Delta FA < 0$), real assets ($\Delta RA < 0$) or cash balances ($\Delta M < 0$).

Clearly there is an infinite number of combinations of changes in liabilities and changes in the holdings of the different types of assets, but overall total assets minus total liabilities (net worth) has declined.

Conversely, a sector that is running a surplus (that is, it is spending less than it is earning or in the parlance used above, it is investing less than it is saving) must be using the surplus funds to:

- Repay debt ($\Delta L < 0$)
- Build up cash balances ($\Delta M > 0$)
- Increase its financial assets (increasing lending) ($\Delta FA > 0$)
- Increase real assets ($\Delta RA > 0$)

More generally, a given sector (for example, household, firm, government) may be running a surplus, but could choose to obtain additional funds by increasing its liabilities by borrowing and incurring debt (ΔL). It can apply those funds to accumulating more financial assets (ΔFA) or building cash balances (ΔM).

Again, there is an infinite number of combinations of changes in liabilities and changes in the holdings of the different types of assets, but overall total assets minus total liabilities (net worth) has increased.

If we wanted to complicate matters we could decompose ΔFA, ΔM and ΔL further, by recognising that a given sector may sell some financial assets and buy new financial assets. Similarly, it might refinance by taking on new liabilities when existing liabilities need to be repaid. So, the entries in Figure 6.7 are to be considered as **net transactions**.

We also have to be cautious in our terminology when considering the different sectors. If we are considering members of the household sector, then it is clear that if they spend less than their income and thus save, they are deferring current consumption in the hope that they will be able to command greater consumption in a future period. The increase in their net worth provides for increased future consumption for the household.

Similarly, for a business firm, if it is spending less than it is earning, we consider it to be retaining earnings, which is a source of funds to the firm in the future.

In summary, net worth for a sector in a given period is the residual after all the uses and sources of funds have been accounted for. From an accounting perspective, net worth is equal to the difference between total assets and total liabilities.

It follows that a change in net worth over the period of interest is equal to the difference between the change in total assets and the change in total liabilities. If total assets increase by more(decrease by less) than total liabilities increase(decrease) then the net worth of the sector has risen.

We consider the private domestic sector as a whole (the sum of the households and firms) to be saving overall if total investment by firms is less than total saving by households. From the national accounts, we consider that households save and firms invest.

However, in the case of the government sector such terminology would be misleading. If the government spends less than it takes out of the non-government sector in the form of taxation we say it is running a fiscal surplus. A fiscal deficit occurs when its spending is greater than its taxation revenue.

But a fiscal surplus does not increase the capacity of the sovereign government to spend in the future in the same way that a surplus (saving) increases the capacity of a household to spend in the future.

REMINDER BOX

As we saw in Chapter 1, a sovereign, currency-issuing government faces no intrinsic financial constraints, and can at any time purchase whatever is for sale in the currency that it issues. Its capacity to do so is not influenced by its past spending and revenue patterns.

Figure 6.8 provides the most comprehensive framework for analysing the flow of funds because it brings together the current transactions (income and expenditure), the financial transactions, and the capital transactions that we have dealt with earlier. The capital and financial transactions are captured in changes to the balance sheet (Figure 6.6).

Figure 6.8 A complete sector uses and sources of funds statement

Uses	Sources
Current expenditure Δ Net worth (saving)	Current receipts
Δ Financial assets (lending) Δ Money (cash balances) Δ Real assets (investment)	Δ Liabilities (borrowing) Δ Net worth (saving)
Σ	Σ

Note that when we talk about the sovereign government we are excluding the levels of government that do not issue the currency. State and local governments are more like households or firms in that respect, although they do have the capacity to tax and issue fines.

The transactions above the dotted line comprise the income statement and record current expenditure (uses). The balancing item above the dashed line constitutes the change in net worth (ΔNW) or 'saving'.

The changes in the balance sheet are shown below the dashed line and the balancing item is once again, the change in net worth (ΔNW).

You can see that we could cancel out the change in net worth (ΔNW) because it is the balancing item in both the income statement and the change in the balance sheet. This would leave us with the accounting statement that sources of funds to a sector through current income and borrowing must be used for current expenditures, investment, lending, and/or building up cash balances.

6.6 The Flow of Funds Matrix

The T accounts tracing the sectoral sources and uses of funds can be summarised for all sectors in the economy by the flow of funds transactions matrix, a stylised version of which is shown in Figure 6.9.

The overriding accounting rule that governs the presentation of the flow of funds accounts is that for the economy as a whole and for each sector in the economy, the total sources of funds must be equal to the total uses of funds. Remember that sources of funds provided by the various sectors in the economy are used by those sectors.

Figure 6.9 (taken from Ritter, 1963) shows three sectors and the total economy. At the most aggregate level, the three sectors could be the private domestic sector, the government sector and the external sector.

For each period being accounted for, the statistician would record the flows of funds that related to each of the row categories in the matrix. Most importantly, we have learned that for every deficit sector, which saves less than it invests, there has to be offsetting surpluses in at least one other sector.

Ritter (1963: 228) described this result as:

an interlocking self-contained system … [which] shows, for a specified time period, the balanced sources-and-uses-of-funds statements for each sector, the interrelations among the sectors, and the aggregate totals of saving, investment, lending, hoarding, and borrowing for the economy as a whole. Any one sector may invest more or less than it saves, or borrow more or less than it lends. However, for the economy as a whole, saving must necessarily equal investment, and borrowing must equal lending plus hoarding.

Thus, a deficit sector – which saves less than it invests – must be offset by at least one other surplus sector to net the flows to zero.

What are the advantages of presenting economic data in this way? Various practical uses can be made of the information provided in the flow of funds accounts.

Figure 6.9 A stylised three-sector flow of funds matrix

	Sector 1		Sector 2		Sector 3		Sector 4	
Flow	**U**	**S**	**U**	**S**	**U**	**S**	**U**	**S**
Saving (ΔNW)								
Investment (ΔRA)								
Lending (ΔFA)								
Cash balances (ΔM)								
Borrowing (ΔL)								

Source: Adapted from Ritter's (1963: 228–9) economy-wide flow of funds matrix.

- They provide information of all financial flows within the economy on a sector-by-sector basis. They allow researchers and policymakers to understand how funds flow from any one sector through the banking system and on to final users by, for example, firms engaged in productive investment.
- They allow researchers and policymakers to monitor major economic trends such as the changing indebtedness of the sectors of the economy and the sources of funding for the respective sectors. For example, an understanding of the flow of funds accounts would have indicated the growing indebtedness of the private sector prior to the GFC, and perhaps alerted policymakers to the likely financial instability arising from it.
- They allow researchers to study saving patterns in the economy, and show us where the savings of a sector are being deployed. The accounts can tell us which sector(s) are accumulating surpluses or deficits and the division between financial and real assets. They also allow us to understand patterns of gross capital formation.
- They permit researchers to examine the dynamics of such concepts as household wealth. We can learn how household balance sheets change over time and how that wealth is composed. For example, one of the hallmarks of the period leading up to the GFC in many countries was the shift in household wealth to riskier categories, such as share purchases sourced from margin loans. The shift in importance in overall wealth from more secure home mortgages to more risky sources of wealth was significant because it exposed the economies to an increased risk of financial instability.
- Central banks use the flow of funds accounts to help them estimate the sensitivity of the economy to changes in the availability of credit.

Flow of funds accounts and the national accounts

The flow of funds accounts complement the national accounts and the balance of payments accounts, which are produced by national statistical agencies on a regular basis as a way of measuring economic activity in total and across the broad economic sectors.

We will consider the balance of payments accounts in Chapter 24.

There are important differences between the flow of funds accounts and the national accounts, which can be summarised as:

- The national accounts contain no data pertaining to financial transactions, borrowing, lending or changes in cash balances. Only non-financial transactions are measured. The flow of funds accounts fill that void.
- The national accounts focus on the current flows of final expenditure, output and income. As we saw in Chapter 4, transactions that involve so-called double counting or intermediate transactions are excluded from the calculations of final expenditures. The flow of funds accounts allow us to trace transactions involving assets that have been created in past periods.
- The structure of the national accounts is such that consumer durable expenditure is included under current expenditure when conceptually it should be considered as investment activity. In the flow of funds accounts all sectors can invest and save.

Conclusion

We began this chapter with a discussion of stocks and flows and then introduced the concept of sectoral balances. We showed how a sector's net income flows are related to its accumulation of financial assets and liabilities, and how one sector's balances are interrelated to another's balances. Finally, we introduced balance sheets used to account for an individual economic entity's uses and sources of funds.

We have seen throughout the chapter how the sectoral balances framework, and the accounting structures that underpin it, empower us to fact-check the internal logic of arguments made by politicians and media commentators about such things as government and private debt retrenchment.

This framework alone does not allow us to determine the validity of politically driven pronouncements such as 'austerity measures will stimulate growth'. At that point, we also need to apply theory but we can – and should – still use the sectoral balances framework to draw macroeconomic inferences about how sectoral balances will respond to the imposition of austerity.

So, if a politician says that the government and non-government sectors should simultaneously reduce their net indebtedness (increase their net wealth by running surpluses) then we know that that is not possible unless the current account surplus grows. In other words, the politician who advocates belt tightening by both the domestic private sector and the government sector in order to reduce indebtedness must also explain their plan for inducing foreigners to increase their own indebtedness to make this possible.[1] We don't have to resort to theory to draw those sorts of conclusions.

References

ABS (Australian Bureau of Statistics) (2014) *Australian System of National Accounts: Concepts, Sources and Methods*, ABS: Canberra.

OECD (2015) Organisation for Economic Co-operation and Development, Economic outlook annex tables. Available at: http://www.oecd.org/eco/outlook/economic-outlook/, accessed 16 September 2018.

Ritter, L.W. (1963) "An Exposition of the Structure of the Flow-of-Funds Accounts", *The Journal of Finance*, 18(2), May, 219–230.

Endnote

1. Such a policy prescription was advocated for Greece in 2012 and beyond in order to reduce both government and private indebtedness. However, the Troika that pushed the policy never admitted that this could not happen unless countries like Germany increased their own indebtedness by running current account deficits against Greece.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER 7 METHODS, TOOLS AND TECHNIQUES

Chapter Outline

Learning Objectives

- Understand the terminology associated with macroeconomic modelling.
- Gain a facility for solving simple linear macroeconomic models.
- Present a macroeconomic model in graphical form.
- Construct and interpret index numbers associated with economic variables.
- Compute growth rates.
- Recognise that economic concepts and relationships can be presented in different ways, such as algebraic and graphical and through a narrative.

7.1 Overview

In macroeconomics we often deploy symbols to represent real world variables of interest, such as real GDP, consumption, and investment. In this case, while the symbols can have an abstract meaning (for example, Y is real GDP) they will also usually have a quantitative analogue (for example, for the September quarter 2013, real GDP in Australia was estimated to be \$A387,031 million).

In Chapter 1, the concept of a model was introduced. A model is a generalisation about the way a system functions or behaves. It could easily be a narrative statement such as "a household will consume a proportion of their income after tax". That theoretical statement might then be examined for its empirical relevance but will also stimulate further theoretical work trying to provide an explanation for that conjectured behaviour.

A macroeconomic model expresses our theoretical conjectures about the relationships between the main macroeconomic variables such as employment, output and inflation.

In economics, like other disciplines that use models, the narrative statement might be simplified with some mathematical statement involving symbols. In this context, the models will be represented by a number of equations (which could vary from one equation to hundreds or even thousands of equations) that describe relationships between the variables of interest. Thus, mathematics is a form of shorthand in terms of concisely representing relationships between variables. We can then apply the basic rules of algebra to conduct our analysis.

We typically use letters (such as Y) to denote different macroeconomic variables. A variable can take on different values in different time periods. The correspondence between the shorthand symbol and the variable is not always intuitive but we maintain consistent conventions throughout this textbook.

Greek symbols (such as α) are often used to denote the parameters of the model that contribute to the formalisation of the relationships between variables. In the first instance these parameters are usually assumed to be constant over time.

So Y is often used to denote real GDP or national income (but it can also be used to denote total output). C is usually used to denote final household consumption and I total private investment. X is typically used to denote exports and M imports, although in some macroeconomic textbooks M is used to denote the stock of money. We use M only to denote imports.

There are several types of variables used in macroeconomic models and the following classification is useful:

- **Endogenous** or dependent variable – its value is determined by the solution to the model and is thus dependent on the values of other variables.
- **Exogenous** or autonomous variable – its value is given in advance of solving the model.

As noted, variables are related by way of equations which express the structure of the macroeconomic model. Usually a variable that we seek to explain is written on the left-hand side of the equals sign (=) and is then expressed in terms of some other variables, placed on the right-hand side of the equals sign, that we consider are influential in explaining the value and movement of the left-hand side variable of interest.

The relationship between the variables on the right- and left-hand sides of the equation is described in terms of some **coefficients** (or parameters).

For example, $y = 2x$ is an equation which says that variable y is equal to two times variable x. The equals sign (=) tells us that the left side of the equals sign is of the same magnitude as the right side (that is, an equation has equal left and right sides).

You solve an equation by substituting values for the unknowns. The number 2 in the equation is called a coefficient and is an estimate of the way in which y is related to x.

So if $x = 1$, then we can solve for the value of $y = 2$ as a result of this equation.

A coefficient can also be called a **parameter**, which is a given in a model and might be estimated using econometric analysis (regression) or assumed by intuition so that the coefficient's value is strictly unknown.

For example, we might have written the above equation as $y = \text{b}x$, where b is the unknown coefficient. You will note that we would be unable to 'solve' for the value of y in this instance even if we knew the value of x. In the case above where we said $x = 1$, then all we could say is that $y = \text{b}$. We would thus need to know what b was before we could fully solve for y.

There are several types of equations that are used in macroeconomic models.

- **Identity equation** – is an expression that is true by definition and is usually based on an accounting statement, as was noted in Chapter 6. For example, we will see that GDP is equal to the sum of the expenditure components, which is true as a result of the way we set up the national accounts and define the expenditure components.
- **Behavioural equation** – captures the hypotheses we form about how a particular variable is determined. These equations thus represent our conjectures (or theory) about how the economy works and obviously different theories will have different behavioural equations in their system of equations (that is, the economic model).
- **Equilibrium equation** – is an expression that captures a relationship between variables that defines a state of rest.

While the example of $y = 2x$ was easy to solve once we knew the value of x, sometimes it is useful to have models which we cannot solve for numerical values of the unknown variables of interest. However, we may be able to simplify the equations to show the structure of the model in terms of what is important to advance our understanding of the relationship between our aggregates.

7.2 Basic Rules of Algebra

You will need to learn some basic algebraic rules that are used to manipulate equations and solve for unknowns of interest. Often you will need to rearrange a given equation in order to determine the solution for the variable of interest.

Model solutions

In a system of equations, the values of some variables are unknown and are only revealed when we **solve** the model for unknowns.

So if we have these two equations, which comprise a **system** of equations:

(7.1) $y = 2x$

(7.2) $x = 4$

then x is a predetermined variable (with the value of 4) and is thus exogenous. You do not know the value of y in advance and you have to solve the equations to reveal its value, so it is endogenous. To solve this system, we substitute the value of x in Equation (7.2) into Equation (7.1) so we get:

$y = 2$ times 4

$y = 8$

More generally the solution of a **system of structural equations** entails expressing each of the endogenous variables, say $y_1, y_2, \ldots, y_n$ as functions of the exogenous variables $x_1, x_2, \ldots, x_m$, so that there are $n + m$ 'solution'

BOX 7.1 RULES OF ALGEBRA

Addition and subtraction

In general, what we add to or subtract from one side of the equation, we have to add to or subtract from the other side to maintain the equality.

Given an equation $y = x$, then we know that the equivalent expression is $y \pm z = x \pm z$. So, for example, $y = x$ is equivalent to $y + 2 = x + 2$.

We can also substitute an expression from one equation into another and maintain equivalence.

For example, we might have $y = 2x$, and $x = 6z$. In this case, it is equivalent to write $y = 2(6z) = 12z$.

Multiplication and division

Given an equation $y = x$, then we know that the equivalent expression is $3y = 3x$ or $y/3 = x/3$. If we multiply or divide the left-hand side of the equation by a variable (or more complex algebraic expression) then we have to multiply or divide the right-hand side of the equation by the same variable (or expression). Dividing by zero is not allowed, however, and multiplying by zero is not very helpful!

equations. In the above example, two solution equations are formed from one endogenous variable (y in Equation (7.1)) and one exogenous variable (x in Equation (7.2)).

Then the known values of the m exogenous variables can be substituted into each of the n equations which yield solutions for the endogenous variables, say $y_1^*, y_2^*, \ldots, y_n^*$. There are constraints on the initial system of (structural) equations for there to be a unique solution for the unknown exogenous xs and endogenous ys. These include the condition that the number of equations equals the number of unknowns, in this case $n + m$.

In modelling an economic system, it often becomes very complicated to ascertain which variables can be considered endogenous and which are truly exogenous. At the extreme, everything might be considered endogenous and then things get mathematically complex. There is a body of econometric theory which explores the problem of identifying values of coefficients in empirical work, but this topic falls well beyond the scope and aims of this textbook.

7.3 A Simple Macroeconomic Model

An example of an identity is the famous national income identity depicting aggregate demand and output, which we considered in Chapter 4:

$$(7.3) \qquad Y \equiv C + I + G + X - M$$

Recall that Y is GDP (total income and output), C is household consumption expenditure, I is private capital formation or investment expenditure, G is total government expenditure, X is total exports and M is total imports. We have used the identity sign (≡) instead of the equals sign (=) to distinguish Equation (7.3) from a behavioural equation which is always expressed using an equals sign.

This identity is also an equilibrium condition in the simple national income model but it provides no information about how the right-hand side variables behave, that is, what factors influence them. To advance that understanding we form theories about the determinants of these variables which are expressed in behavioural equations. An example of a behavioural equation is the simple consumption function:

$$(7.4) \qquad C = C_0 + cY_d$$

which says that final household consumption (C) is equal to a constant (C_0) plus a proportion (c) of final disposable income (Y_d). The constant component (C_0) is the consumption that occurs if there is no income and can be considered to be dissaving.

Note that subscripts are often used to add information to a variable. So we append the subscript d to our income symbol Y to qualify it and denote disposable income (total income after taxes).

We also use subscripts to denote time periods when we are considering a variable over time. So Y_t indicates we are considering the value of Y at time period t. Similarly, Y_{t-1} refers to the value of Y at time period $t - 1$, where the lag (the -1) depends on the periodicity of the data. If we were using quarterly data, then $t - 1$ would be previous quarter and so on.

In macroeconomics, some behavioural coefficients are considered important and are given special attention. So the coefficient c in the consumption function is called the marginal propensity to consume (*MPC*) and denotes the extra consumption per dollar of extra disposable income. So if $c = 0.8$ we know that for every extra dollar of disposable income that the economy produces, 80 cents will be spent on consumption.

The *MPC* is intrinsically related to the marginal propensity to save (*MPS*) which is the amount of every extra dollar of disposable income that is saved after households decide on their consumption. So $MPS = 1 - MPC$ by definition.

The importance of *MPC* is that it is one of the key determinants of the expenditure multiplier (see Section 7.5). We will consider this in Chapter 15 when we discuss the expenditure multiplier.

We have already introduced the distinction between exogenous variables and endogenous variables which are determined by solving the system of equations.

An exogenous variable is known in advance of 'solving' the system of equations. We take its value as given or predetermined. We might say, by way of simplification, that government spending (G) in Equation (7.3) is equal to \$100 billion, which means that its value is known and not determined within the model.

The identity and behavioural equations form a macroeconomic system. This is of course a very simple system. For the sake of exposition, we might assume that the economy is closed, which means there are no exports or imports. In that case, the national income identity becomes $Y = C + I + G$.

We also assume that there is no taxation in the model, so that disposable income is equal to total income (Y).

The model is now:

$$(7.5) \qquad Y = C + I + G$$

$$(7.6) \qquad C = C_0 + cY$$

For simplicity, we will assume that I and G are exogenous because their values are known in advance. The remaining two variables Y and C are endogenous so their values are dependent on the solution to the model.

How do we solve for Y?

We substitute (7.6) for C in (7.5) such that:

$$(7.7) \qquad Y = C_0 + cY + I + G$$

We can now rearrange this (noting we have terms in Y on both sides) by subtracting cY from both sides as per our algebraic rules. This gives:

$$(7.8) \qquad Y - cY = C_0 + I + G$$

You will note that there are only predetermined variables (knowns) on the right-hand side now.

Because Y is what we call a common factor on the left-hand side we can write $Y - cY$ as $Y(1 - c)$.

$$(7.9) \qquad Y(1 - c) = C_0 + I + G$$

Note that this has not affected the terms on the right-hand side.

We now divide both sides by $(1 - c)$, which maintains the equality between the two sides of the equation. This isolates (or solves for) Y on the left-hand side. Thus:

$$(7.10) \qquad Y = [C_0 + I + G]/(1 - c)$$

So in words, the equilibrium value of Y is a function of the autonomous variables in the model $C_0 + I + G$.

We call Equation (7.10) the **reduced form solution** of the model in which there are only exogenous or predetermined variables on the right side and an unknown variable on the left-hand side of the equation.

In a macroeconomic model, all the endogenous variables can be expressed in reduced form. So in our example, our solution for consumption would be:

$$(7.11) \qquad C = C_0 + cY = C_0 + c[C_0 + I + G]/(1 - c) = [C_0 + c(I + G)]/(1 - c)$$

Make sure you can derive the steps that we would take to obtain this solution.

The reduced form of the system allows us to conduct sensitivity analysis, which involves changing the values of the exogenous variables or the coefficients (in this case, the *MPC*) and analysing the impact on the endogenous variables, C and Y, in the model.

As an example, what would be the impact of an expansion in government spending G on national income Y? Note that we assume the other exogenous variables are unchanged.

From Equation (7.10), we know that when $G = G_0$:

$$(7.10a) \quad Y_0 = [C_0 + I + G_0]/(1 - c) = [C_0 + I]/(1 - c) + [G_0]/(1 - c)$$

Note that we have separated G from I and C_0.

If G rises to G_1, then:

$$(7.10b) \quad Y_1 = [C_0 + I]/(1 - c) + [G_1]/(1 - c)$$

So, a change in Y of $(Y_1 - Y_0)$ is the difference between (7.10a) and (7.10b):

$$(7.10c) \quad (Y_1 - Y_0) = [G_1 - G_0]/(1 - c)$$

To simplify our notation, we will usually denote the change in a variable using the Greek symbol Δ. So Equation (7.10c) would be written as:

$$(7.10d) \quad \Delta Y = \Delta G/(1 - c)$$

where the time span denoted by Δ is revealed by the context.

By dividing both sides of the equation by ΔG, we can express Equation (7.10d) as:

$$(7.10e) \quad \Delta Y/\Delta G = 1/(1 - c)$$

The right-hand side of Equation (7.10e) is known as a **multiplier** because it tells us the magnitude of the change in Y for a unit change in G. We will examine multipliers in more detail in Chapter 15, where we include taxation and exports and imports in an open economy.

7.4 Graphical Depiction of a Macroeconomic Model

We will also express our theories in graphical terms, which are an alternative to mathematical representation.

1. When income is zero, household consumption is positive. Household consumption rises proportionately with disposable income but the proportion is less than one.
2. $C = C_0 + cY_d$, where $0 < c < 1$ and C_0 is a constant (fixed value). The less than sign (<) tells us that the c (MPC) lies between the value of zero and one, that is, it is positive but less than one.

   For now, we assume taxes are zero which means that national income (Y) and national disposable income (Y_d) are equal.

Considering the consumption function, if $C_0 = 100$, and $c = 0.8$, and $Y = 400$, we can solve the equation $C = C_0 + cY$ by inserting the known values of the parameters and the explanatory variable (in this case, income) into the equation and solving it. Thus:

$$(7.12) \quad C = C_0 + cY = 100 + 0.8 \times 400 = 420$$

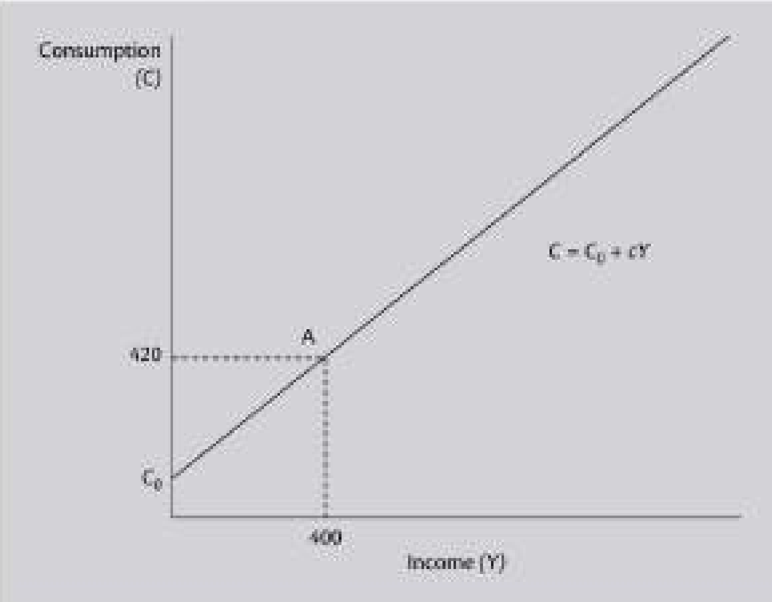
Figure 7.1 shows the consumption function for Equation (7.12). You can see that by tracing a vertical line from where disposable income equals 400 up to the graph line (A) and then a horizontal line to the vertical axis we derive the value of consumption by where that line crosses the vertical axis.

The slope of the line is the marginal propensity to consume (c). In Chapter 15 we will deal with applications of the slope of a line when we study the principle of the expenditure multiplier.

To advance our understanding of graphical methods, assume that national income rose to 1,000.

Figure 7.2 shows the combination of Y and C that we have already determined. Point A represents the combination $(Y_1, C_1) = (400, 420)$. It also shows the second combination, denoted B, which is found by the same

Figure 7.1 Consumption function



process as point A. Trace a vertical line from where disposable income equals 1,000 up to the graph line and then a horizontal line to the vertical axis to give the value of consumption. Point B then represents the combination $(Y_2, C_2) = (1{,}000, 900)$.

We can check our graphical solution by calculation using the equation $C = C_0 + cY = 100 + 0.8 \times 1{,}000 = 900$.

We represent the slope of the line that intersects A and B using the following formula:

$$C = \frac{C_2 - C_1}{Y_2 - Y_1} \tag{7.13}$$

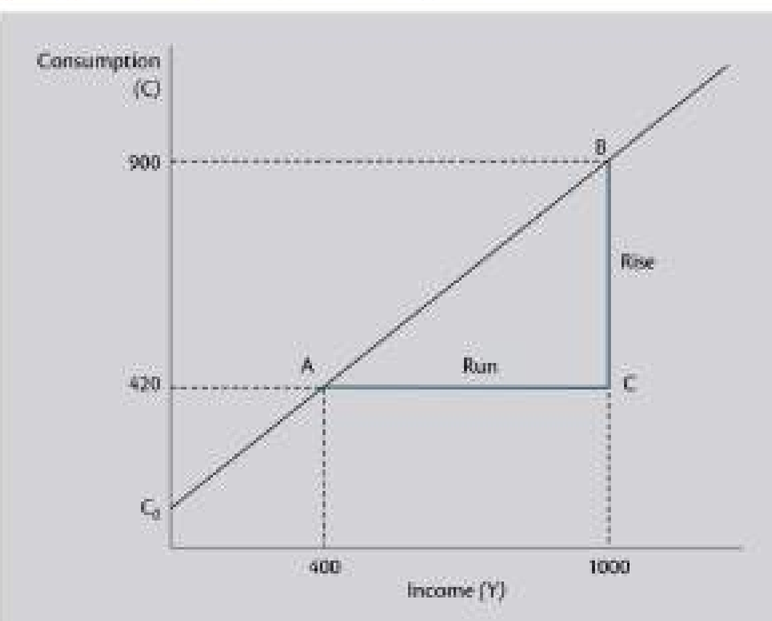
The slope of the line is given by the change in C divided by the change in Y. You can see from Figure 7.2 that this is also the ratio of the segments, BC/AC, or in words **Rise over Run**. This rule generalises to any linear function. Note that for a line with negative slope, Rise would be Fall.

You can check the slope of our consumption function by noting that BC = 480 and AC = 600 so BC/AC = 0.8, which is the MPC (c).

We noted earlier that we usually denote the change in a variable using the Greek symbol Δ. In that notation the slope of the consumption function would be $c = \Delta C/\Delta Y$ and that slope would be constant at all points on the (linear) function, irrespective of the magnitude of ΔY (and the corresponding value of ΔC).

The graphical approach to determining the slope of the line is confined to linear equations. If the function is non-linear (for example, a curve), the slope formula (Equation 7.13) only provides us with the average slope between two points.

Figure 7.2 Slope of consumption function



A more general approach is required when the relationship of interest is non-linear and the slope is constantly varying. In this situation, differential calculus is used and the slope of the function at some specific point is defined by the **derivative** of the function.

The principle of calculating the slope from $c = \Delta C/\Delta Y$ is the same as described above. However, when deploying differential calculus, we are assuming that ΔY is infinitesimally small (that is, a value approaching zero), as is ΔC. Otherwise, in contrast to a linear function, there will be ambiguity in measuring the slope because it will depend on the chosen magnitude of ΔY, as well as the value of Y at which it is measured. Typically, we will be concerned with changes that are not infinitesimally small and so the tools of calculus will be of limited use to us in this textbook. Thus, we will use calculus sparingly but it is useful to understand the basic concept of a derivative.

The slope of a function $y = f(x)$ is $\Delta y/\Delta x$. We can see that $\Delta y = f(x + \Delta x) - f(x)$. Now what would happen if Δx was close to zero?

Take a specific example of a non-linear function, $y = x^2$.

$$\text{If } f(x) = x^2, f(x + \Delta x) = (x + \Delta x)^2 = x^2 + 2x\,\Delta x + (\Delta x)^2$$

Note that we do not expect students to be able to derive these expressions themselves. Rather, we hope you can appreciate the concept being outlined.

Now if we substitute this into the slope formula we get:

$$\frac{\Delta y}{\Delta x} = \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

$$(7.14) \qquad = \frac{x^2 + 2x\Delta x + (\Delta x)^2 - x^2}{\Delta x} \qquad \text{(cancels out } x^2 \text{ and } -x^2\text{)}$$

$$= \frac{2x\Delta x + (\Delta x)^2}{\Delta x} \qquad \text{(divide by } \Delta x\text{)}$$

$$= 2x + \Delta x$$

The derivative of a function $y = f(x)$ is the limiting value of $\Delta y/\Delta x$ as $\Delta x \to 0$.

So from Equation (7.14), the derivative of $y = x^2$ is $2x$, since $\Delta x \to 0$ and can be ignored. Thus the derivative of a non-linear function of x depends on the chosen value of x. The derivative of a function $y = f(x)$ is written as dy/dx or $f'(x)$.

What is meant by the derivative of a function? It means that if $y = x^2$, the rate of change of the value of the function, y, at any point is $2x$. If $x = 4$, then the rate of change in y is 8. If $x = 5$, then the rate of change in y is 10.

More generally, if $y = x^n$, it can be shown that $dy/dx = nx^{n-1}$. Thus, if $y = x^3$, $dy/dx = 3x^2$, which can be evaluated at any value of x.

For the consumption function, $C = C_0 + cY$, applying the general formula for a derivative would yield $dC/dY = cY^{1-1} = cY^0 = c$, which is consistent with what we have already found. Note that $Y^0 = 1$.

7.5 Power Series Algebra and the Expenditure Multiplier

We have introduced the concept of the expenditure multiplier, which we consider further in Chapter 15. It allows us to calculate the total change in national income following a change in one of the autonomous components of aggregate demand, such as government spending or private investment.

Multipliers play a central role in discussions of the economic impact of policy interventions. The multiplier depends on the magnitude of the initial injection of expenditure plus the induced expenditure that follows. We will learn that when, for example, the government increases its spending, national income rises by that amount, which in turn induces further consumption expenditure.

Thus, if one dollar was injected into the economy, through additional spending, total income would initially rise by one dollar. If the marginal propensity to consume was 0.8, then this initial rise in income would induce a rise in consumption of $0.8 \times \$1$ or 80 cents in period 1. This initial \$0.80 rise in induced spending would further induce a rise in consumption in period 2 of 0.8×0.8 or 64 cents and so on. Note that this is a simple, mathematical exposition of what in the real world would be a complex process of adjustment of the economy to an increase of government spending.

The multiplier is thus related to the slope of the function and the algebraic notion of a geometric or power series.

Consider a one dollar increase in government spending with c representing the marginal propensity to consume and n being the number of periods. Then we could write the multiplier over n periods as $k(n)$, where:

$$(7.15) \qquad k(n) = 1 + c + c^2 + c^3 + c^4 + \ldots + c^n$$

The right-hand expression is called a **power series** with each term being a constant multiple of the previous term.

Note that multiplying both sides of (7.15) by c gives:

$$(7.16) \qquad ck(n) = c + c^2 + c^3 + c^4 + c^5 + \ldots + c^n + c^{n+1}$$

If we subtract (7.16) from (7.15) we get:

$$(7.17) \qquad k(n) - ck(n) = k(n)(1 - c) = 1 - c^{n+1}$$

which gives:

$$(7.18) \qquad k(n) = (1 - c^{n+1})/(1 - c)$$

For a large number of periods we consider $n \rightarrow \infty$ so the value of c^{n+1} tends to zero because $0 < c < 1$. We denote the summation by k^*. Hence:

$$(7.19) \qquad k^* = 1/(1 - c)$$

Equation (7.19) shows that the multiplier, k^*, is $1/(1 - c)$, where c is the marginal propensity to consume, when n is a large number. It should be interpreted as the overall impact of the multiplier process, rather than considering the impact over a finite number of periods.

7.6 Index Numbers

An index number allows the comparison between the values of a variable over time. For example in Chapter 4, the construction of a consumer price index (CPI) was described. However, if two or more variables are expressed in index number form, then straightforward comparisons of their respective rates of change (growth rates) can be made.

The creation of an index number requires a starting point (the base period or value) which is usually set at 100. Each observation is then expressed as a percentage of the base year. For example, consider the data in Table 7.1, which shows full-time, part-time and total employment for Australia from 2000 to 2012. The data are reported in units of thousands.

Visual inspection of the data can provide information as to the evolution of employment over this time period in Australia. But what if we wanted to know whether full-time employment had grown more quickly or more slowly than part-time employment since 2000? This is where index numbers are useful.

To convert this data into index numbers we first define the base as the year 2000 and set that index value to 100. We then generate index values of full-time employment for 2001 through to 2012, by comparing employment in each of these years with employment in 2000 (for which the index value is 100). To do this we divide each value of full-time employment by its value in 2000 (6,614.6) and then multiply by 100. The index number for full-time employment in 2001 would thus be 100 × (6,597.5/6,614.6) = 99.7 and so on.

Table 7.2 shows the index numbers corresponding to the three employment time series. Note that while the raw employment data is expressed in units of thousands, the index numbers are unit free. Thus defining a common base year for the three series (that is, values in 2000 are set at 100) means that comparisons between them can be readily made, even though their corresponding absolute values differ quite significantly.

Table 7.1 Employment in Australia, 2000–12

	Full-time employment (000s)	Part-time employment (000s)	Total employment (000s)
2000	6,614.6	2,375.1	8,989.7
2001	6,597.5	2,492.4	9,089.9
2002	6,648.7	2,622.0	9,270.7
2003	6,772.7	27.12.2	9,485.0
2004	6,930.4	2,730.9	9,661.3
2005	7,148.7	2,849.4	9,998.1
2006	7,326.9	2,930.2	10,257.1
2007	7,583.7	2,993.3	10,577.0
2008	7,782.5	3,093.1	10,875.6
2009	7,724.3	3,229.5	10,953.7
2010	7,852.5	3,337.6	11,190.0
2011	8,014.5	3,376.1	11,390.6
2012	8,098.3	3,417.1	11,515.4

Source: Data from Australian Bureau of Statistics, Labour Force.

Table 7.2 Employment indices for Australia, 2000–12

	Numbers in employment			Employment index		
	Full-time employment (000s)	Part-time employment (000s)	Total employment (000s)	Full-time employment (2000=100)	Part-time employment (2000=100)	Total employment (2000=100)
2000	6,614.6	2,375.1	8,989.7	100.0	100.0	100.0
2001	6,597.5	2,492.4	9,089.9	99.7	104.9	101.1
2002	6,648.7	2,622.0	9,270.7	100.5	110.4	103.1
2003	6,772.7	27.12.2	9,485.0	102.4	114.2	105.5
2004	6,930.4	2,730.9	9,661.3	104.8	115.0	107.5
2005	7,148.7	2,849.4	9,998.1	108.1	120.0	111.2
2006	7,326.9	2,930.2	10,257.1	110.8	123.4	114.1
2007	7,583.7	2,993.3	10,577.0	114.7	126.0	117.7
2008	7,782.5	3,093.1	10,875.6	117.7	130.2	121.0
2009	7,724.3	3,229.5	10,953.7	116.8	136.0	121.8
2010	7,852.5	3,337.6	11,190.0	118.7	140.5	124.5
2011	8,014.5	3,376.1	11,390.6	121.2	142.1	126.7
2012	8,098.3	3,417.1	11,515.4	122.4	143.9	128.1

We can see that the index number for full-time employment rose from 100 in 2000 to 122.4 in 2012, a 22.4 per cent increase. In the same period the part-time employment index number rose from 100 in 2000 to 143.9 in 2012, a 43.9 per cent rise which is almost twice as large as the increase in full-time employment.

Note that for any pair of index number observations we can also compute percentage changes. For example, what was the growth of full-time employment in 2008 and 2009?

We know that full-time employment was 7,782.5 in 2008 and had fallen to 7,724.3 in 2009. We could calculate a simple percentage 100 × (7,724.3 − 7,782.5)/7,724.3 = −0.75 per cent. We could find the same result using the index numbers 117.7 in 2008 and 116.8 in 2009 and computing 100 × (116.8 − 117.7)/117.7 = −0.76 per cent. (The small difference here is accounted for by rounding up the index numbers to one figure after the decimal point.)

It is also useful to graph index numbers if we are interested in a visual comparison of the behaviour of different variables.

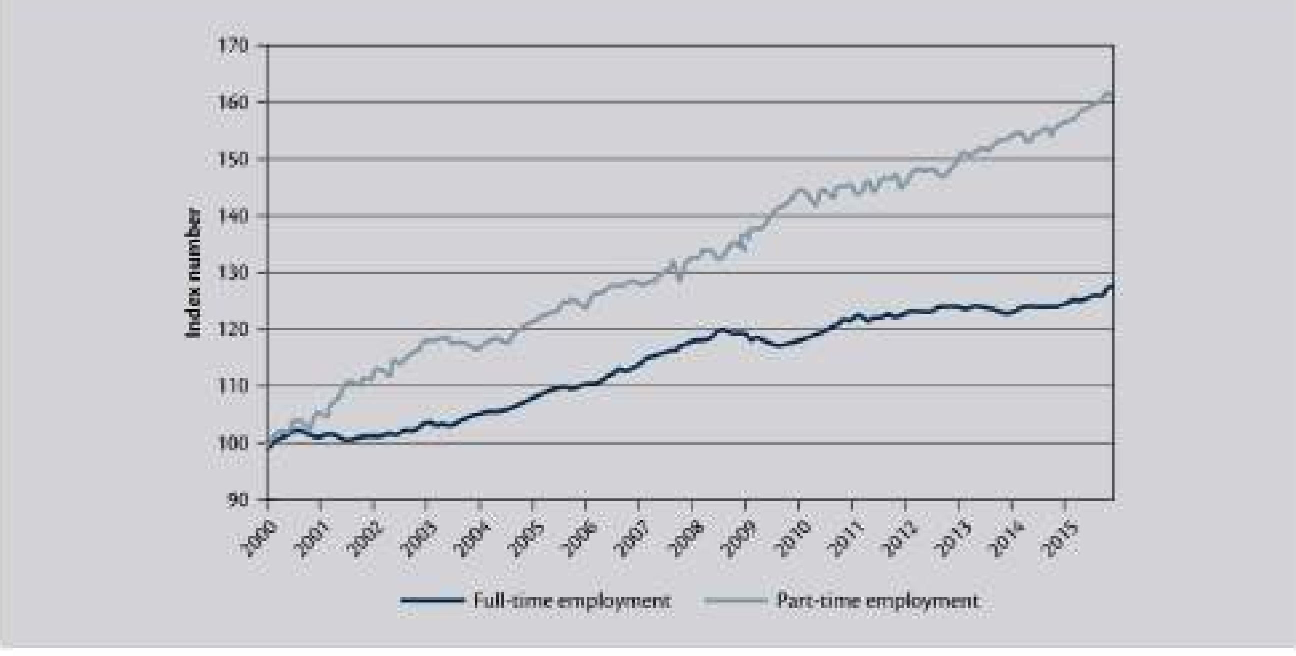
For example, Figure 7.3 allows one to see easily that in 2008, when the financial crisis emerged, full-time employment in Australia contracted while part-time employment accelerated. This observation would motivate a researcher to investigate the labour market processes that underpinned this outcome.

As mentioned above, index numbers also allow us to compare the evolution of two or more variables over time when the underlying units of measurement of each variable are different. For example, for decades real wages in most nations grew in line with labour productivity, which created the space for wages to grow without invoking inflationary pressures. This observation means that economists are often interested in examining the relationship between the two variables over time.

Figure 7.4 shows the relationship between real wages and productivity growth in Australia from 1978 to 2015 in index number form. The base period is March 1982 and the underlying data are quarterly. Productivity is measured in terms of units of real GDP per hour worked while real wages are computed as the nominal wage series ($ per hour) deflated by the consumer price index.

By converting the different series to common index numbers we are readily able to see the comparative behaviour of these related time series variables.

Figure 7.3 Employment index numbers, Australia, 2000–15, January 2000=100



Source: Authors' own. Data from Australian Bureau of Statistics, National Accounts.

Figure 7.4 Real wages and productivity, 1978–2015, March quarter, 1982=100

Source: Authors' own. Data from Australian Bureau of Statistics, National Accounts.

7.7 Annual Average Growth Rates

In Chapter 4 we computed the percentage change in the CPI from one period to the next. We were essentially solving the equation:

$$CPI_t = CPI_{t-1} \times (1 + r^*/100)$$

where r^* is the rate of change of the CPI, expressed as a percentage.

Economists are often required to compute how fast the economy (or some other aggregate) is growing on average over a number of years. We are then calculating a constant compound growth rate.

7.8 Textbook Policy Regarding Formalism

We recognise that students have different ways in which they learn and develop their understanding of concepts. Some prefer the mathematical approach while others prefer the graphical approach. Others still learn better through reading the written word, even though that form of communication is prone to interpretative issues.

All the essential material in the text will therefore be presented in all three ways. Sometimes the mathematical treatment will appear in the Appendix of the relevant chapter (usually the more difficult, advanced material), and sometimes within the main body of the text.

We always aim to promote understanding and believe that a student is entitled to learn in the way that best suits their own proclivities. However, it is also the case that professional economists use a variety of methods including numerical, graphical, algebraic and narrative in their work, and we believe it is important to expose students to the broad range of presentation methods deployed in the real world.

TRY IT YOURSELF

In 1960, real GDP in Australia was $249,083 million and it had grown to $1,508,267 million by 2012. To calculate the average annual compound growth rate over this 52-year period, we need to utilise some simple algebra and deploy the notion of a compound growth rate.

We can write:

$$Y_t = Y_0(1 + r)^t$$

where Y_t is real GDP in 2012; Y_0 is real GDP in 1960; r is the average compound growth rate, expressed as a decimal, that is, $r = r^*/100$; and t is the number of periods over which we are compounding, in this case 52 years.

Here we need to use the natural logarithm (ln) to simplify our task.

The task is to solve for the unknown r:

$$(Y_t/Y_0) = (1 + r)^t$$

$$ln(Y_t/Y_0) = t \times ln(1 + r)$$

$$\frac{ln(Y_t/Y_0)}{t} = ln(1 + r)$$

$$exp\left[\frac{ln(Y_t/Y_0)}{t}\right] = 1 + r$$

$$r = exp\left[\frac{ln(Y_t/Y_0)}{t}\right] - 1$$

Table 7.3 shows the steps in the calculation and you can paste the data into a spreadsheet to derive the same results. The calculations show that the annual average compound growth real GDP growth rate for Australia between 1960 and 2012 was 3.52 per cent.

You can use the formula for any period or data frequency (for example, month, quarter, year) by substituting the appropriate information into the calculation.

Table 7.3 Compound growth rate calculations

Real GDP 1960 $m	Y_0	249,083
Real GDP 2012 $m	Y_t	1,508,267
Time periods (years)	t	52
GDP2012/GDP1960 (Y_t/Y_0)	$= (1 + r)^t$	6.0553
Take log	$= t\ln(1 + r)$	1.8009
Divide by 52	$= \ln(1 + r)$	0.0346
Take antilog	$= 1 + r$	1.0352
	r	0.0352
	r per cent	3.5240

Conclusion

This chapter provided a review of the basic mathematics that is often used in macroeconomics, including algebra, power series, index numbers, and growth rates. A simple macroeconomic model was presented to provide an example of the methodology often used. Our policy regarding formalism in this textbook was clarified: we use three methods to present all the essential material: words, graphs, and mathematics. Each student can choose the method best suited to their learning style.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

8 THE USE OF FRAMING AND LANGUAGE IN MACROECONOMICS

Chapter Outline

Learning Objectives

- Recognise the importance of framing and language in learning macroeconomics.
- Understand the way humans use metaphors for cognition.
- Develop an appreciation for the difficulty involved in changing framing.
- Relate the learning objectives to the pedagogy of Modern Monetary Theory (MMT) and the exposition in this textbook.

8.1 Introduction

Words matter. From the time you were a baby, your parents and other guardians told you stories designed to teach you how to behave. *The Little Engine that Could* taught the value of perseverance; *Yertle the Turtle* instilled in you compassion and willingness to stand up against tyranny; *Grimm's Fairy Tales* warned of the danger of straying too far from parental protection.

Words and lessons you learned in childhood stick with you. They are imprinted in your brain and can be easily triggered. If someone utters the word 'witch', this will likely trigger an old memory of one of the stories you were told as a child. Sometimes you cannot get these memories out of your mind.

George Lakoff, an expert on linguistics, provides the following example. He tells readers 'Don't think of an elephant'. Try as you might, you will think of an elephant. If we go on to say, 'Whatever you do, do not think of *Dumbo*', most of you, especially those raised in Western countries, will think of the character from the popular children's book and Disney movie. Those big floppy ears stick in your mind and make it hard to think of anything else.

There is more to this than simply associating a word (elephant) with a vision (the picture of Dumbo that you recall from childhood). Modern science teaches us that humans think using stories, many of them learned long ago as children. In the thinking process (cognition), you adapt these old stories to new situations to make sense of the world.

Advertisers make use of the modern science of cognition to invoke associations between products they are trying to sell and neural networks that have already been implanted in your brain. For example, a toilet bowl cleaner is pitched by associating it with a young, smiling mother who cares for her cute, happy toddler. Politicians also use science to trigger both good and bad associations to win votes. For example, a politician who is running on a 'tough on crime' platform will obliquely refer to a notorious case to trigger fear in the minds of voters. American President Ronald Reagan famously told a story about a 'welfare queen' mother who drives a 'Cadillac' during his campaign to associate welfare programmes with cheaters and wasteful spending.[1]

Economists unavoidably use stories to communicate their beliefs, although they often do not realise that the way they frame their argument can promote or hinder the lesson they are trying to teach. Nor do they always understand what neural networks the words they use will trigger in the minds of the audience. For example, the word 'debt' is necessarily a loaded term with associations of imprudence, burden, and default. Once the neural networks associated with these stories are fired up, it is extremely difficult to avoid putting negative connotations onto the word – and thus the concept of – 'debt'.

In general, conservative economists have been much more aware of the science of cognition, and have successfully used it to pervasively embed their ideology into economic discourse, including in the classroom and in the policy-making sphere. Unfortunately, progressive economists have generally adopted the conservatives' terminology and framing, undercutting their attempt to provide an alternative to the dominant, conservative, approach to economics.

In this textbook, we strive to use language that is consistent with our progressive approach to economics. In this chapter, we will discuss how language and framing is important to teaching the Modern Monetary Theory (MMT) alternative.

8.2 MMT and Public Discourse

This textbook provides an integrated exposition of MMT, which stands in contradistinction to the mainstream macroeconomics that is taught in many universities and is dominant in policy-making circles. In this chapter, we discuss the way in which framing and language are used to advance conceptual ideas and how people can be persuaded to believe propositions that have no foundation in reality.[2] Learning the importance of framing and language is a precondition for gaining a fundamental grasp on what is knowledge and what is not.

Macroeconomics concepts, such as real GDP, inflation, the unemployment rate, the fiscal deficit, and the interest rate, make headlines on a daily basis. Finance segments on national news broadcasts increasingly discuss macroeconomic issues with frequent recourse to complex terms that are not well understood by most media commentators or the public.

Consequently, from the outset the public discourse contains significant errors that render it almost impossible for participants to make informed assessments of macroeconomic developments independently of the associated politics.

Economic concepts such as the government fiscal deficit contain nuances that make an unambiguous assessment of their meaning difficult. One cannot conclude for example, that a deficit equal to two per cent of GDP signals a more expansionary fiscal stance by government than a deficit of half the size. These complexities are lost to the public but are fundamental to an accurate understanding of the issues.

Social media exacerbates this problem. The so-called blogosphere is populated with self-styled macroeconomics experts whose claims about such things as the state of government finances are erroneous at best. Appeals to 'common sense', and reliance on logic or intuition have all become dangerous norms in these forums, reinforcing the argument that not all opinion should be given equal privilege in public discourse. Our propensity to generalise from personal experience, as if that experience equates to general knowledge, continually leads these 'Facebook' macroeconomics debates into combative dead ends of false reasoning from false propositions.

The problems inherent in communicating complex economic concepts and evidence are amplified by the ideological assumptions that currently dominate public debate. Economics as an academic discipline has come to be defined by a set of beliefs that are associated with the dominant free market paradigm. The consequence of this is a narrow debate that excludes the lessons of history and the possibility of alternative economic paradigms that offer realistic insights into current economic conditions and related policy options.

Conservative think tanks and media outlets produce an array of 'research' or 'policy' reports such that the public understanding has become straitjacketed by orthodox concepts and conclusions that in themselves are erroneous, but also lead to policy outcomes that undermine prosperity and subvert public purpose. The public's willingness to tolerate mass unemployment, rising income inequality and poverty is a manifestation of this syndrome.

Prior to the Global Financial Crisis (GFC), mainstream economists pronounced the business cycle dead and declared that we had entered a period of "great moderation" (Stock and Watson, 2002; Lucas, 2003; Bernanke, 2004). These economists categorically failed to foresee the catastrophic consequences of the policy of deregulation of the labour and financial markets that they promoted.

It is reasonable to expect that professional failure on the scale of the GFC would have led to a re-evaluation of the paradigm within which these economists work, and to major changes in economic curricula and research. Mainstream economists however have failed to alter their teaching approaches in any significant way.

The fact that mainstream macroeconomics has retained its hegemonic status in the face of its failure to resonate with reality is due in no small measure to the way economics debates are framed in the public discourse. Framing refers to the way an argument is conceptualised and communicated by speakers and listeners. Conceptual models and representations are used in this process. Research in cognitive philosophy and cognitive linguistics suggests that the models that constrain our thinking operate at a largely unconscious level, and that the abstract concepts we draw on are 'largely metaphorical', 'imaginative', and 'emotionally engaged' (Lakoff and Johnson, 1999: 3–4).

Proponents of neoclassical macroeconomics have been extremely successful in their use of common metaphors to advance their ideological interests. Ideologically loaded myths – in modern parlance, 'fake knowledge' – are accepted unchallenged by the public as truths. Thus ideology triumphs over evidence and we accept falsehoods as truth.

Recent psychological studies have highlighted the extent to which pre-existing biases influence the way in which we interpret factual information, including straightforward statistical data. This presents a further problem for the communication of research findings that may be counterintuitive or controversial, or that challenge the dominant discourse about public policy design, as in the case of economic austerity or climate change.

Further, it is often the case that the language that is used to advocate a progressive alternative actually serves to reinforce conservative myths (see Lakoff, 2010; Wray, 2013). It has been suggested that language is so important in learning that there is a need to develop new terms that are not associated with orthodox economics in order to communicate a new progressive macroeconomics, such as is developed in this textbook. But there is a tension between any benefits that may accrue from the development of such a new language, and the necessity to couch the argument in terminology that the public are familiar with when discussing economic outcomes. There remains a role for better education of essential concepts in addition to a reconsideration of how the concepts are related.

In this textbook, we base our exposition around MMT, which is a coherent macroeconomic narrative providing new insights into the policy opportunities available to a government and their likely outcomes.

MMT has struggled to gain traction in wider economic and political debates due to:

1. An incomplete understanding of key macroeconomic terms amongst economics commentators, especially journalists, and the wider community, including policy makers and the general public; and

2. The deployment of key macroeconomic terms in the context of pervasive cultural metaphors to support policy interventions that effectively benefit a privileged few at the expense of the majority.

In this chapter, we have already provided a brief conceptual basis for understanding how the language we use constrains our thinking. We will go on to examine some of the key metaphors used to reinforce the flawed message of orthodox economics.

8.3 Two Visions of the Economy

Shenker-Osorio (2012) juxtaposes two visions of the economy. Figure 8.1 represents the conservative mainstream economics view, where the basic assumption is that "people and nature exist primarily to serve the economy" (Shenker-Osorio, 2012: Location 439).

This narrative tells us that a competitive, self-regulating economy will deliver maximum wealth and income if left free from government intervention. The economy is enshrined as an entity – almost a deity – that is essentially beyond our control although it recognises our endeavours, and rewards – or punishes – us accordingly. We are required to have faith, work hard and make the necessary sacrifices for the good of the economy; those who do not are rightfully deprived of such rewards.

The economy is also constructed as a living entity. If the government intervenes in the competitive process and provides an avenue by which the 'undeserving' (lazy and so on) can also receive rewards, then the system becomes 'sick'. The solution is to restore the economy's natural processes (its 'health'), which entails the elimination of government intervention such as minimum wages, job protection, and income support.

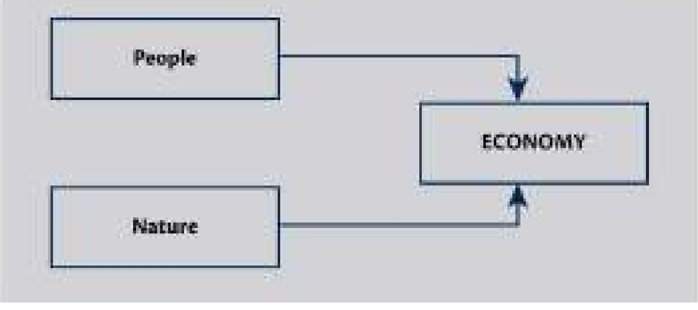
The key messages are "self-governing and natural", which force the obvious conclusion that "government 'intrusion' does more harm than good, and we just have to accept current economic hardship" (Shenker-Osorio, 2012: Location 386).

Although subscribers to this view would have us believe that this is a rational narrative, in fact it represents a type of 'magical thinking' more appropriately associated with medieval views on the relationship between individuals and the world.

The narrative teaches us that our own success or failure is due primarily to our own efforts and these are considered to be independent from the success of the system. The extent to which private wealth can be seen as linked to socioeconomic status and stable, high-quality infrastructure is minimised. Similarly, the unemployed are seen as being responsible for their jobless status, when in reality a systemic shortage of jobs explains their plight.

This narrative is so powerful that progressive politicians and commentators have become seduced into offering 'fairer' alternatives to the mainstream solutions rather than challenging assumptions behind them root and branch. For example, progressives timidly advocate more gradual fiscal austerity when they should be comprehensively rejecting it on the basis of evidence that it fails, and advocating larger deficits to solve the massive rates of labour underutilisation that burden most economies.

Figure 8.1 The conservative economic construction

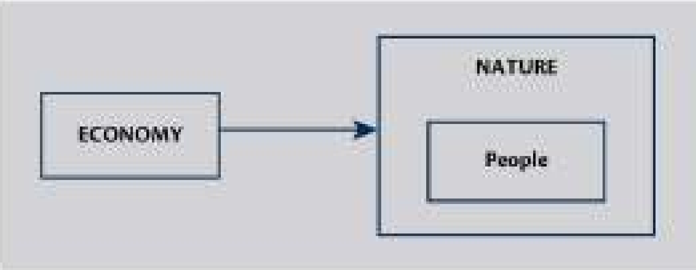


Source: Authors' own. Derived from discussion within Shenker-Osorio (2012).

Progressives and conservatives are hostage to the same erroneous beliefs about the way the economy operates and the public is compelled to believe there is no alternative (TINA) to the damaging economic policies being introduced.

Figure 8.2 represents an alternative view of the economy, where the economy works for us as our construction and people are organically embedded within and nurtured by the natural environment. Accordingly, the economy is seen as a constructed object and policy interventions around it should only be appraised

Figure 8.2 The economy is us



Source: Authors' own. Derived from discussion within Shenker-Osorio (2012).

in terms of reaching our broad goals, which a progressive vision would articulate in terms of advancing public well-being and maximising the potential for all citizens within the limits of environmental sustainability. The focus then shifts to one of placing our human goals at the centre of our thinking about the economy, which echoes the principles of functional finance (Lerner, 1943). Consistent with this, MMT highlights the irrelevance of a narrow focus on the fiscal balance without reference to a broader human context.

In this narrative, people create the economy. There is nothing 'natural' about it. Concepts such as a 'natural rate of unemployment', which imply that it should be left to its own equilibrating forces to reach its natural state, are erroneous. Governments can always choose and sustain a particular unemployment rate. We created government as our agent to do things that we cannot easily do ourselves and we understand that the economy will only serve our common purposes if it too is subjected to active oversight and control (Mitchell and Muysken, 2008).

The two visions of the economy can be summarised in value terms as individualistic (Figure 8.1) and collectivist (Figure 8.2). For a progressive economist, collective will is important because it provides the political justification for more equally sharing the costs and benefits of economic activity. Progressives have historically argued that government has an obligation to create work if the private market fails to create enough employment. Thus, government is empowered to spend as much as it takes to ensure there are enough jobs available for all those who want to work.

The contest between the two visions outlined in Figures 8.1 and 8.2 has spawned a long-running academic debate. It played out during the Great Depression, which taught us that policy intervention is vital in order to rein in the chaotic and damaging forces of greed and power that underpin the capitalist monetary system.

We learned that so-called 'market signals' do not deliver satisfactory levels of employment and that the system can easily come to rest with mass unemployment. We learned that this malaise was brought about by a lack of spending and that the government had the spending capacity to redress these shortfalls and to ensure all those who wanted to work could do so.

From the Great Depression and responses to it, we learned that the economy is a construct, not a deity, and that we could control it through fiscal and monetary policy in order to create desirable collective outcomes. We understood that the economy is our creation, designed to deliver collective benefits, not an abstract entity that distributes rewards or punishments according to some moral framework. The government is not a moral arbiter but a functional entity serving our needs. People's experiences during this period led to a complete rejection of neoclassical macroeconomics by Keynes and others (see Chapters 11 and 12).

The OPEC oil price hikes in the 1970s provided the switch point that saw the conservative ideas represented in Figure 8.1 regain ascendancy, despite their ideas having been cast into disrepute during the 1930s.

The resurgence of the free market paradigm since that time has been accompanied by a well-crafted public campaign, where framing and metaphor triumph over operational reality or theoretical superiority. This process has been assisted by business and other anti-government interests, which have provided significant funds to emergent conservative, free market 'think tanks'. Beder argues that these institutions fine-tuned "the art of 'directed conclusions', tailoring their studies to suit their clients or donors" (1999: 30). Politicians parade these so-called 'independent' research findings as the authority needed to justify their deregulation agendas, and the distortion to the policy debate is reinforced by the erroneous claims emanating from the think tanks.

8.4 Cognitive Frames and Economic Commentary

An overview of the assumptions that underpin cognitive linguistics and cognitive philosophy can help to explain why it has been difficult to mount a coherent challenge to orthodox macroeconomic views. Lakoff and Johnson (1999: 3) make three basic claims:

1. The mind is inherently embodied.
2. Thought is mostly unconscious.
3. Abstract concepts are largely metaphorical.

We can summarise what this means in the following way. Embodiment relates to everyday activities such as grasping something with our hands, or standing and walking. We relate these physical actions via the use of metaphors to try to understand the world around us. For example, we understand concepts associated with 'more' or 'less' in terms of a direction 'up' or 'down' as a consequence of our physical experience of interacting with our environment. Metaphors concerned with direction are also employed in our reasoning about being happy or sad (as in 'that was an uplifting film' or 'I'm feeling down') and success or failure (as in 'they are climbing the corporate ladder' or 'they have fallen from grace'). We think about quantity in terms of direction such that 'more' is conceptualised as an upward trend, and 'less' is figured as downwards. In relation to our general thinking about money, 'more' is typically good and 'less' is typically bad. The idea that a bigger fiscal deficit can be a desirable thing is thus counterintuitive in a very basic sense.

The development of an economic pedagogy and an understanding of the policy opportunities arising from a more complete understanding of macroeconomic terms need to be communicated in a way that exploits our capacity to construe the same information in different ways.

8.5 Dominant Metaphors in Economic Commentary

The mainstream theoretical models in macroeconomics that reinforce the Figure 8.1 perspective do not embrace the basic characteristics of the real world economy and consistently perform poorly when evaluated against their predictive accuracy of real world events.

We have already noted that the dominance of the mainstream macroeconomics narrative in the public domain is achieved through a series of linked myths that are reinforced with strong metaphors. Table 8.1 shows some of the popular examples of metaphors used by mainstream economists and commentators to focus their attack on government spending, deficits, public debt and income support payments for the most disadvantaged workers. As we learned in Section 8.4, each of these metaphors is designed to reinforce the main core values that the mainstream paradigm seeks to promote, such as self-discipline, independence, ambition, wealth, and sacrifice (see Figure 8.1). These metaphors obscure the truth and lead us to support policies that make us worse off even when there are alternatives that would make us all, in a collective sense, better off.

8.6 Face to Face: Mainstream Macro and MMT

The implicit metaphors buried in terms such as those in Table 8.1 reinforce a series of propositions promoted by mainstream economists to focus their attack on government intervention. In the remaining chapters we provide an alternative theoretical approach to reveal the fallacies inherent in mainstream macroeconomics. In the remaining sections of this chapter we summarise the relevant operational reality as presented by MMT, which we will develop in detail throughout the rest of this textbook (see Mosler, 1997–8; Mitchell, 1998; Wray, 1998; Mitchell and Muysken, 2008). As you will see, these counter the metaphors presented in Table 8.1.

Table 8.1 Examples of neoclassical macroeconomic metaphors

Focus of attack	Metaphorical claim	Implied meaning
Government spending	The nation is living beyond its means	Excess spending requires sacrifice Cuts needed immediately
	Nation has maxed out its credit card	Run out of money due to irresponsible spending
	Spending like a drunken sailor	Wanton irresponsibility and delinquent behaviour
Fiscal balance	Budget black hole	Budget beyond human control like the collapse of massive star
	Deteriorating state of the budget	Budget is like a body and is in state of ill health requiring emergency surgery. There is no Alternative (TINA)
	Mushrooming budget deficit	Budget is an organic entity that has grown out of control
	The nation has run out of money, it is broke	Government budget is like a household budget; the economy is like us
Public debt	The nation is bankrupt	Nation is a badly managed insolvent firm
	The public debt mountain	Debt is dangerous and insurmountable
	Burdening our grandchildren	Debt threatens fundamental unit of society (the family) and undermines future prosperity
	Mortgaging the future	Current government debt compromises future spending
Income support	Welfare dependency	Welfare net is like a drug for the populace, encouraging ill health and addiction
	Dole bludgers, Skivers	Unemployed people are lazy and undeserving

Mainstream Fallacy 1: The government faces the same 'budget' constraint as a household

Many of the widely used metaphors, like those in Table 8.1, seek to directly equate government spending with that of a reckless *individual* householder (splurging on credit card debt, neglecting to pay the mortgage, failing to care for their children). The household budget analogy is false. Households use the currency and must finance their spending. A sovereign government issues the currency and must spend before it can subsequently tax or borrow. A currency-issuing government can never be revenue-constrained in a technical sense and can sustain deficits indefinitely without solvency risk.

Implications

- The household budget analogy is inapplicable to a currency-issuing government.
- Our own personal budget experience generates no relevant insights when we analyse government's outlays and revenues.
- An alternative narrative must highlight the special characteristics of the government's currency monopoly.

Mainstream Fallacy 2: Fiscal deficits(surpluses) are bad(good)

Fiscal deficits are neither good nor bad. In accounting terms, they equal the non-government surplus. In behavioural terms, they are required when the spending intentions of the non-government sector are insufficient to ensure full utilisation of the available productive resources. The context matters because the fiscal outcome is a vehicle to achieving socioeconomic goals rather than an end in itself.

Similarly, fiscal surpluses are neither good nor bad and may be harmful in some circumstances. For a nation with strong net exports, quality public services, and national income levels sufficient to support the private sectors' saving desires, a surplus may be required to contain nominal aggregate demand and avoid inflation.

There are two problems with using the term 'fiscal deficit'. First, while many speak as if a fiscal deficit is a policy choice made by the government, in fact the fiscal outcome is determined by the state of overall activity, and is largely beyond the control of government. If private spending is weak, then the deficit will typically rise as tax revenue declines. Movements in the fiscal balance are thus ambiguous. A given fiscal outcome can result either because the government has taken discretionary fiscal policy decisions to maintain full employment given the spending and saving decisions of the non-government sector; or because non-government spending has fallen and the automatic stabilisers have led to a decline in tax revenue net of transfers and an increase in unemployment. When private spending collapses and the deficit rises, the correct response is to increase discretionary net public spending, not cut it. Government does not budget for a fiscal deficit. Instead, whether a deficit results (and how government ought to react to the budgetary outcome) depends on the performance of the economy.

Second, the term has a negative connotation because a deficit signifies a shortfall, which while accurate in an accounting sense, is highly misleading in the context of the positive contribution that a deficit makes to the net financial wealth of the non-government sector. Government deficits are the sole source of net financial assets for the non-government sector. All transactions between agents in the non-government sector net to zero. This accounting reality means that if the non-government sector wants to net save in the currency of issue, then the government has to be in deficit.

The sectoral balances derived from the national accounts generalise this result and show that the government deficit(surplus) *always* equals the non-government surplus(deficit). Fiscal surpluses destroy private wealth by forcing the private sector to liquidate its wealth to get cash and destroy liquidity (debiting reserve accounts), which is deflationary. With an external deficit, fiscal surpluses result in increasing private domestic sector debt levels and cannot represent a sustainable long-term growth strategy. Ultimately, the decision by the private domestic sector to increase its net saving and reduce its debt levels will interact with the fiscal drag coming from the surpluses and force the economy into recession. The cyclical component of the budget will then push the deficit back into a (bad) deficit.

Implications

- Understanding the context in which a particular fiscal outcome occurs is crucial to a reasoned assessment of the appropriateness of the fiscal position.
- The fact that a fiscal deficit allows the non-government sector to save overall and a fiscal surplus destroys non-government wealth should be understood and promoted.

Mainstream Fallacy 3: Fiscal surpluses contribute to national saving

A currency-issuing government does not save in its own currency. Fiscal surpluses do not represent 'public saving' that can be used to fund future public expenditure. Saving is an act of forgoing current spending to enhance future spending possibilities and applies to a financially constrained non-government entity, such as a household.

Fiscal surpluses provide no greater capacity to governments to meet future financial needs, nor do fiscal deficits erode that capacity. The constraints on government spending are not financial, but are defined by the accessibility of real resources that are available for sale in the currency that the government issues (see Fallacy 7).

Implications

- A currency-issuing government never needs prior funds in order to spend and thus never needs to 'save'.
- Fiscal surpluses(deficits) destroy(augment) non-government financial wealth.

Mainstream Fallacy 4: The fiscal outcome should be balanced over the economic cycle

Recognising that the fiscal outcome is endogenous means that the government cannot realistically target a particular outcome because changes in private spending, for example, can thwart any efforts made by the government to achieve that target. The responsible strategy for a government is to allow its fiscal balance to adjust to the level of net spending required to achieve full employment given the spending decisions of the non-government sector, irrespective of the state of the business cycle.

The national accounts tell us that for a nation with an external deficit, a 'balanced budget rule' is tantamount to requiring the private domestic sector to record a deficit of equal magnitude to the external deficit. This is unlikely to be a sustainable strategy.

Further, a countercyclical fiscal strategy does not require the government to achieve a surplus. The concept of counter cyclicality more correctly refers to the direction of change rather than the level of the fiscal outcome. The government should not increase its discretionary net spending if the economy is already at full capacity and it is satisfied with the private spending mix. Such an expansion would be procyclical. But a fiscal stance that results in a steady deficit is desirable where the external deficit is steady and the private domestic sector is saving overall.

Implications

- Fiscal rules defined in terms of public debt or deficit ratios are unlikely to be consistent with the responsible management of fiscal policy.
- A currency-issuing government should pursue functional targets such as full employment and allow its fiscal outcome to adjust accordingly.

Mainstream Fallacy 5: Fiscal deficits drive up interest rates and crowd out private investment because they compete for scarce private saving

This is a specific version of Fallacy 2. While currency-issuing governments do not need to sell bonds, the fact that they do creates no competition for finite savings between public and private borrowers. First, government deficits stimulate growth and private savings as national income grows. Second, the funds that the government absorbs when it sells bonds to the non-government sector ultimately originate from the net financial assets created by past deficits. Third, bank lending is not reserve constrained and banks will extend loans to any creditworthy customer. If the banks are short of reserves, then they borrow from each other in the interbank market. Ultimately, they can always borrow from the central bank. For that reason, selling government bonds to banks does not reduce the capacity of banks to make loans to customers.[3]

Moreover, deficits place downward pressure on interest rates. Debt issuance serves to allow the central bank to maintain a positive target interest rate by providing investors with an interest bearing asset that drains the excess reserves in the banking system resulting from deficit spending. If these reserves were not drained, then in an environment of government deficits, the overnight interest rate would fall (due to competition by banks to rid themselves of the non-profitable reserves) and this might compromise the central bank's target interest rate unless it offers a return on excess reserves, which most do.

Implications

- Central banks use public debt as part of an interest rate targeting strategy.
- Public debt does not fund government spending.
- Currency-issuing governments do not need to borrow.

Mainstream Fallacy 6: Fiscal deficits mean higher taxes in the future

Taxes serve many purposes (reducing private sector purchasing power, reducing consumption of harmful goods and services such as tobacco, and so on) but none of these purposes relate to funding government spending.

In a fiat monetary system, where the currency has no intrinsic worth, the government needs to transfer real goods and services from the non-government to the government sector to facilitate its economic and social programme. In this context, a primary function of taxation is to promote offers from private individuals to government of goods and services in return for the necessary funds to extinguish tax liabilities. The crucial point is that the funds necessary to pay the tax liabilities are provided to the non-government sector by government spending.

Accordingly, government spending, if sufficient, provides the paid work, which eliminates the unemployment created by the taxes.

By depriving the non-government sector of purchasing power, taxes attenuate aggregate demand so that government can create non-inflationary, real resource space to accommodate public spending.

Importantly, every generation can freely choose the level of taxation it pays because it determines, through the political process, the size of government and the real resource space it will utilise. Past fiscal deficits never need to be paid back by the current generation, or by any future generation.

Implications

- Progressives should refer to 'public' money rather than 'taxpayers' money.
- Taxpayers do not fund government spending.
- Taxes are required to reduce the capacity of the non-government sector to command real goods and services so that the government can utilise them.

Mainstream Fallacy 7: The government will run out of fiscal space (or money) if it overspends

This is related to Fallacies 1, 5 and 6. Conservative politicians and economic commentators often claim the government will run out of money if it does not curb spending. They attempt to give this statement authority by appealing to our intuition and experience, via the household budget analogy, to claim that governments, like households, have to live within their means. This analogy resonates strongly with voters because it is easily relatable: we intuitively understand that as individuals, we cannot live beyond our means indefinitely.

A currency-issuing government has no intrinsic financial constraints. Government will never run out of money to build a hospital or pay health professionals, but the materials to build the facility and the skilled workers to run it may not be available. Fiscal space is thus more accurately defined as the real goods and services available for sale in the currency of issue. These are the 'means' available to government to fulfil its socioeconomic charter. The currency-issuing government can always purchase whatever is for sale in its own currency.

This fallacy is also related to the intergenerational (ageing) population claims that pension and healthcare systems will be unsustainable in the future. There are no financial constraints to stop a currency-issuing government providing first class healthcare and/or pensions in the future. The challenge of rising dependency ratios will be whether productivity growth ensures there are adequate real goods and services available to maintain growth in living standards with fewer workers available. These are not financial constraints.

Another, related, claim is that the sovereign issuer of currency is at risk of default if the public debt ratio rises above some threshold (often construed as 80 per cent). As long as the government only issues debt in its own currency and provides no assurances about convertibility into another currency, the default risk is zero.

Implications

- Fiscal space is not defined in terms of some given financial ratios (such as a public debt ratio).
- Fiscal space refers to the extent of the available real resources that the government is able to utilise in pursuit of its socioeconomic programme.
- There is no default risk for debt issued by a government in its own currency.

Mainstream Fallacy 8: Government spending is inflationary

All spending (private or public) is inflationary if it drives nominal aggregate demand above the real capacity of the economy to absorb it. Increased government spending is not inflationary if there are idle real resources that can be brought back into productive use (for example, unemployment).

Related fallacies include the claims that issuing bonds to the central bank – the so-called 'printing money' option – devalues the currency, whereas issuing bonds to the private sector reduces the inflation risk of deficits. Neither claim is true. First, there is no difference in the inflation risk attached to a particular level of net public spending when the government matches its deficit with bond issuance relative to a situation where it issues no bonds. Bond purchases reflect portfolio decisions regarding how private wealth is held. If the funds that we used for bond purchases were spent on goods and services as an alternative, then the fiscal deficit would be lower as a result. Second, the provision of credit by the central bank (in return for treasury bonds) will only be inflationary if there is no fiscal space (see Fallacy 7).

Hyperinflation examples such as 1920s Germany and Zimbabwe in the early 2000s do not support the claim that deficits cause inflation. In both cases, there were major reductions in the supply capacity of the economy prior to the inflation episode.

Implications

- When the economy is at full employment, all spending carries an inflation risk and issuing treasury bonds does not mitigate the risk associated with public spending.
- Government spending should seek to bring idle resources back into productive use.
- The limits of non-government public spending are defined by the fiscal space available, which in turn is a function of the availability of idle resources.

Mainstream Fallacy 9: Fiscal deficits lead to big government

Fiscal deficits may reflect any size of government. Even small governments will need to run continuous deficits if there is a desire within the non-government sector to save overall and the policy aim is to maintain full employment levels of national income.

Economic theory does not specify an optimal size of government. The call for smaller government reflects a purely ideological stance with no basis in economic theory. The size of government will reflect the preferences of the population for public provision of goods and services and infrastructure.

Implications

- The size of government is a political choice rather than an economic necessity.
- Even small governments will typically run continuous deficits to maintain full employment.

8.7 Framing a Macroeconomics Narrative

While MMT provides an internally consistent and empirically robust account of the way the economy works, it also clearly challenges the way in which humans think about macroeconomic concepts. But through careful use of language and avoiding the established macroeconomics metaphors, we can develop a coherent understanding of the conceptual framework developed in this textbook.

Language and metaphor examples

The primary metaphor **Purposes Are Destinations** relates to the subjective judgement that we want to achieve purposes and we succeed when we reach a destination.

MMT defines the destination in terms of people rather than in terms of an independent 'economy'. For example, we might define our goal as full employment, or a 'zero waste of people'. This destination is reached not when the public debt ratio is *x* per cent, but when all those who want to work can find a job; for example, when measured unemployment is less than two per cent and there is zero underemployment.[4] That is the desired specific destination and allows us to adopt a positive focus to our analysis. If we become caught up debating financial ratios such as the size of fiscal deficits and the like, we merely reinforce the conservative frames about sacrifice, solvency, and rectitude (Table 8.1).

A key issue then concerns language and terminology. As Shenker-Osorio notes, the term 'spending' may be problematic because to 'spend' means to 'use up'. This implies that what is spent was finite, has now gone and is no longer with us. This language supports erroneous assumptions about what a government that is the sovereign issuer of its own currency does with its money (Shenker-Osorio, 2012: Location 452).

While MMT teaches us the reality that government spending injects net financial assets into the system to meet "critical human needs" (Shenker-Osorio, 2012: Location 452), our language has to reinforce rather than undermine that understanding.

We might then say that 'government spending puts money in our pockets' while government taxing 'takes money out of our pockets'. This is not only more accurate, but it also provides a different connotation to both government spending and government deficits (that is, to put more money into our pockets than taken out through taxes).

Still, 'spending' retains that negative connotation even when applied to private sector spending. While we like the benefits of our spending (a nice restaurant dinner, for example), we do not like 'spending' our money. Instead of focusing attention on the *act* of spending, we need to focus on the *results* of government's spending. Rather than saying 'governments spend', we might instead say:

Governments invest to enhance our purposes.

Invest is a term we readily relate to wealth creation, that is, 'building up' not 'using up'. This invokes the **More is Up** metaphor, which reverses the current negative connotations of spending.

Similarly, the term 'budget deficit' has two negative connotations. First, using 'budget' triggers the fallacious household analogy. Second, 'deficit' signifies a shortfall, and a failure. Throughout this textbook we use the term 'fiscal deficit' and eschew the descriptor 'budget' to avoid invoking the household analogy.

The term 'deficit' is clearly accurate in an accounting sense, but highly misleading in the context of the contribution that a fiscal deficit makes to the net financial wealth of the non-government sector. But trying to replace the term deficit would almost certainly result in a total loss of meaning, further complicating the pedagogy.

We thus do not invent a new term for deficit, but rather seek to exploit the **More is Up** metaphor to advantage by always relating the government deficit to its non-government manifestation, namely the rise in net financial assets that a deficit provides. So we would say:

The government deficit rose and generated higher levels of wealth for households and firms.

This type of language also more aggressively invokes what is known as the **Event-Structure** metaphor, which positions "purposes as destinations to be reached" (Lakoff and Johnson, 1999: 94). The destination must be prominent in the narrative and then we must specify the causal chains through which the purposes are achieved. Causation is linked to concepts of forced movement where, for example, we might say that the government deficit is the application of a force (the injection of net financial assets), which causes a change of state (higher income, employment, or wealth).

Lakoff and Johnson (1999: 91) say the concept of forced movement suggests to us that the "movement would not have occurred without the application of the force" (that is, the economy is a construct rather than a natural self-equilibrating system beyond our control) and that the improvement in national income or employment was the direct result of the cause (the government stimulus).

People respond to the logic where **Causes are Forces** and **Causation is Forced Movement**. Concepts that are framed in this way are more easily learned and understood. Under the 'economy as a natural system or deity' paradigm (Figure 8.1), government regulations and interventions are unnecessary and damaging. In the MMT alternative, the same actions force movement towards our desired destinations. Without the force, the current state does not change and the goals are thus not achieved.

We want government to act on our behalf to move us from State A (less desirable) to State B (closer to our purpose). It is important to note that the economy has no goals. They are *our* goals and *we* use, manage, and control the economy to achieve our goals.

Further, careful choice of frames to promote in the public debate is important. For example, while conservatives focus on concepts such as 'tax relief' to attack government, progressives should frame the debate by emphasising the need for 'unemployment relief', a frame that makes unemployment painful and a reduction in the pain a desired goal.

Relatedly, we avoid debating within the frames that conservatives use. For example, attacking the implementation of fiscal austerity as being 'too fast' still implies (concedes) that the desirable alternative is a more gradual (managed) arrival at the same goal – a reduction in the government deficit.

The more productive frame would be to explicate the functional role of government deficits and continually reference the human costs of conservative policies. A demand for 'relief' from high unemployment reframes the debate and focuses on the failure by conservative forces to reach the desired destination (access to employment). Every statement must reinforce the purposes and destinations that define how we feel about policy.

Fiscal space

In Chapter 22, we outline the concept of fiscal space. National accounting conventions show that the government deficit *always* equals the non-government surplus, and the government surplus *always* equals the non-government deficit.

As we learned in Chapter 6, the combination of external deficits and a desire by the private domestic sector to save overall means that the non-government sector will act as a drain on overall spending in relation to income flows. This means a continuous fiscal deficit is required to sustain a given level of economic activity.

It is thus a *normal* state for a nation to have continuous government deficits over each economic cycle, rising and falling with fluctuations in non-government sector net spending. Rarely will a government surplus be appropriate.

Further, government deficits are not just appropriate in times of recession or slow growth. They are required whenever there is a non-government desire for a surplus, which is the typical case. We might thus continually reinforce the frame:

Government deficits are normal, surpluses are atypical.

This means that 'balanced budgets over the cycle'-type arguments are destructive and fall into the misleading 'deficits are bad' frame.

We thus frame the concept of fiscal space in terms of the idle real resources that can be brought into productive use via higher government spending and/or lower taxation. The idle resources signal that the government deficit is too low or the surplus is too large. The desired destination is zero waste and the required action is a larger deficit.

Costs of a public programme

The emphasis on real resource availability as the demarcation of fiscal space is also related to a frame linked to cost. We often hear or read statements such as:

- The costs to the taxpayer of programme A are huge;
- The nation cannot afford the cost of that programme.

If we were to take a public employment programme that required government to spend \$*x* billion in wages, capital equipment, administration and oversight, we might reasonably ask about the cost of that programme. The conservative frame tells us that the cost is \$*x* (the figure that appears in the annual fiscal documents against the programme).

An MMT frame considers the \$*x* in the fiscal papers to be of little interest. The actual cost of the programme is the change it causes in the usage of **real resources**: more consumption by the formerly unemployed workers who now have jobs, some equipment and project supervision costs, and so on. An additional cost would be the opportunity costs of such a programme, which are minimal given that the unemployed are idle. In fact, in this frame, the increased use of the real resources provides benefits to both the individuals and society, so the use of the term 'cost' would be misleading.

When we ask whether the nation can afford a policy initiative, we should ignore the \$*x* and consider what real resources are available and the potential costs of putting unemployed resources to work (which would be nearly zero) plus the costs of shifting already-employed workers to the government's programme (opportunity cost). The available real resources constitute the fiscal space. The fiscal space should then always be related to the purposes to which we aspire, and the destination we wish to reach.

The MMT alternative framing

In Table 8.2 we present a summary of the MMT alternatives to the framing that was presented in Table 8.1.

Comparison of Table 8.2 with Table 8.1 will make clear how different the two approaches are. We can make use of our understanding of linguistics to provide a framing consistent with the policy space available to sovereign nations that issue their own currency.

Table 8.2 Examples of MMT macroeconomic metaphors

Focus of attention	Metaphorical claim	Implied meaning
Government spending	Government spending puts money in our pockets.	Government spending increases non-government's income.
	Government invests in the productive capacity of our nation.	Government spending increases our capacity to provide for the needs of our population.
	Government cannot run out of money; the true constraint is our nations' resources.	There is no financial constraint although there could be a resource constraint.
Fiscal balance	Government deficits allow us to save.	Government deficits equal non-government surpluses.
	Government surpluses equal non-governmental deficits.	Government surpluses reduce non-government financial saving.
	The fiscal balance is largely determined by the economy's performance.	Fiscal balance outcome is not discretionary.
	The nation cannot run out of its own money.	Government 'budget' is NOT like a household budget.
Public debt	The government's debt is our asset.	Government debt provides a risk-free financial asset to strengthen non-government portfolios.
		Government debt helps to stabilise the financial system.
Income support	We take care of our own.	As our government faces no financial constraints, it uses its fiscal capacity to ensure resources are mobilised to care for its population.
	A good nation supports its people.	Unemployment is always evidence of a policy failure, a failure to put resources to work.
	A good nation helps to support those in other nations who need help.	Rich nations help to mobilise resources needed by poorer nations.

Conclusion

MMT provides an accurate description of how the monetary system actually operates.

The decision to pursue a policy framework consistent with Figure 8.1 (individualistic) or Figure 8.2 (collectivist) is purely ideological. The orthodox economic rhetoric associated with Figure 8.1 trades on an incomplete understanding of macroeconomic realities and exploits powerful metaphors to ensure that these realities and related policy opportunities are obscured from vision.

References

Beder, S. (1999) "The Intellectual Sorcery of Think Tanks", *Arena Magazine*, 41, June/July, 30–2.

Bernanke, B.S. (2004) "The Great Moderation", Speech to the Eastern Economic Association, Washington DC, 20 February.

Connors, L.J and Mitchell, W.F. (2017) "Framing Modern Monetary Theory", *Journal of Post Keynesian Economics*, 40(2), 239–59.

Lakoff, G. (2010) "Untellable Truths", *Huffington Post*, 10 December. Available at: http://www.huffingtonpost.com/george-lakoff/untellable-truths_b_794832.html, accessed 26 April 2017.

Lakoff, G. and Johnson, M. (1999) *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*, New York: Basic Books.

Lerner, A. (1943) "Functional Finance and the Federal Debt", *Social Research*, 10(1), 38–51.

Lucas, R.E. Jr. (2003) "Macroeconomic Priorities", *The American Economic Review*, 93(1), March, 1–14.

Mitchell, W.F. (1998) "The Buffer Stock Employment Model – Full Employment without a NAIRU", *Journal of Economic Issues*, 32(2), 547–55.

Mitchell, W.F. and Muysken, J. (2008) *Full Employment Abandoned: Shifting Sands and Policy Failures*, Aldershot: Edward Elgar.

Mosler, W. (1997–8) "Full Employment and Price Stability", *Journal of Post Keynesian Economics*, 20(2), Winter: 167–182.

Shenker-Osorio, A. (2012) *Don't Buy It: The Trouble with Talking Nonsense about the Economy*, New York: Public Affairs, E-book references.

Stock, J. and Watson, M. (2002) "Has the Business Cycle Changed and Why?", *NBER Macroeconomics Annual*, MIT Press, 159–218.

Wray, L.R. (1998) *Understanding Modern Money: The Key to Full Employment and Price Stability*, Cheltenham: Edward Elgar.

Wray, L.R. (2013) "A Meme for Money", Working Paper No. 736, Levy Economics Institute.

Endnotes

1. This was in the early 1980s and voters at the time would also have understood that there was an implicit racial dimension to US President Reagan's story. While in fact most recipients of 'welfare' (the main programme was Aid to Families with Dependent Children) were white, the common perception of 'welfare moms' was that they were disproportionately African Americans. Further, the Cadillac was an American-made luxury car favoured by African Americans while rich white Americans were increasingly choosing foreign-made luxury cars (particularly those made in Europe).
2. See Connors and Mitchell (2017) for a detailed discussion. This chapter borrows heavily from that article.
3. Indeed, because government bonds are very safe assets, they are held in bank portfolios to serve as collateral against their own borrowing. This makes it easier for them to borrow reserves from other banks when they run short. If anything then, holding such bonds might actually increase bank willingness to lend to customers.
4. Full employment can be reached before the measured unemployment rate is zero because some of those between jobs prefer to spend some time seeking an appropriate position; this transition between the condition of unemployment and employment will result in a non-zero measured unemployment rate. Full employment has often been defined as 'more vacancies than job seekers', in other words, a situation in which there are plenty of job openings for those looking for work.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

PART B

CURRENCY, MONEY AND BANKING

CHAPTER

9 INTRODUCTION TO SOVEREIGN CURRENCY: THE GOVERNMENT AND ITS MONEY

Chapter Outline

Learning Objectives

- Explain why a fiat currency is valued and is acceptable in domestic transactions.
- Recognise the distinction between fixed and floating exchange rate regimes and their significance for the conduct of macroeconomic policy.
- Understand how IOUs are created and extinguished.

9.1 Introduction

In this chapter we will examine in more detail several of the concepts briefly introduced in earlier chapters of this textbook. We first turn to the money of account and the nation's currency and note that the latter is no longer backed by a precious metal, such as gold. We argue that the so-called **fiat currency** is valued and widely used in transactions because it is required as the means to pay taxes and other obligations levied by the state. All financial stocks and flows are denominated in the national money of account. In this context the financial system can be viewed as a record of transactions, that is, a scoreboard. We then examine the difference between floating and fixed exchange rate systems.

Government and non-government IOUs are denominated in the national currency, or money of account. After defining leveraging (use of debt), we argue that these different types of IOUs can be conceived of as a financial pyramid, with government IOUs at the top.

Finally, we emphasise the need to use the term 'money' very carefully to avoid confusion.

9.2 The National Currency (Unit of Account)

Let us look at money as the **unit of account** in which stocks and flows are denominated.

One nation, one currency

In Chapter 6, we introduced the concept of the money of account. The Australian dollar, the US dollar, the Japanese yen, the British pound, and the European euro are all examples of a money of account. The first four are each associated with a single nation which, throughout history, has been the usual situation: 'one nation, one currency'. There have been a few exceptions to this rule, including the modern euro, which is the money of account adopted by several countries that have joined the Economic and Monetary Union of the European Union (EMU). When we address the exceptional cases, such as the EMU, we will carefully identify the differences that arise when a currency is used, but not issued, by a nation.

Most of the discussion that follows will be focused on the more common case in which a nation adopts its own money of account. The government of the nation issues a currency (usually consisting of metal coins and paper notes of various denominations) denominated in its money of account. Spending by the government as well as tax liabilities, fees, and fines owed to the government are denominated in the same money of account. These payments are enforceable by law. More generally, broad use of a nation's money of account is ensured by enforcing monetary contracts in a court of law, such as the payment of wages.

In many nations there are private contracts that are written in foreign monies of account. For example, in some Latin American countries it is common to write some kinds of contracts in terms of the US dollar. It is also common in many nations to use US currency in payment. Many contracts governing international trade are denominated in US dollars, even if neither party uses that specific currency as their own. According to some estimates, the total value of US paper currency circulating outside America exceeds the value of US paper currency used at home. Much of this is thought to be involved in illegal activities, including the drug trade.

Thus, one or more foreign monies of account as well as the corresponding foreign currencies might be used in addition to the domestic money of account and the domestic currency denominated in that unit. Sometimes this is explicitly recognised, and permitted by, the authorities, while other times it is part of the underground economy that tries to avoid detection by using foreign currency. While we recognise these deviations from the 'one currency, one nation' rule, they generally account for a relatively small proportion of transactions and contracts, most of which will be denominated in the nation's own money of account.

Sovereignty and the currency

The national currency is often referred to as a **sovereign currency**, that is, the currency issued by the sovereign government. The sovereign government retains a variety of powers for itself that are not given to private individuals or institutions. Here, we are only concerned with those powers associated with money. The sovereign government alone has the power to determine which money of account it will recognise for official accounts. Further, modern sovereign governments alone are invested with the power to issue the currency denominated in each nation's money of account. For example, if any entity other than the US government tried to issue US currency it would be prosecuted as a counterfeiter, with severe penalties being imposed. (Enemy nations do sometimes try to undermine a nation's economy by counterfeiting its currency, hoping to cause inflation and destroy trust in the currency.)

As noted above, the sovereign government imposes tax liabilities (as well as fines and fees) in its money of account, and decides how these liabilities can be paid: that is, it decides what it will accept in payment so that taxpayers can fulfil their obligations.

Finally, the sovereign government also decides how it will make its own payments, when it purchases goods or services, or meets its own obligations, such as pensions to retirees. Most modern sovereign governments make payments in their own currency, and require tax payments in the same currency. For reasons that we will examine, requiring tax payments in the governments' currency ensures that the same currency will be accepted in payments made by government.

What 'backs up' the currency?

There has been ongoing confusion surrounding sovereign currency. For example, many policy makers and economists have had trouble understanding why the private sector would accept currency issued by the government when it made purchases. Some have argued that it is necessary to 'back up' a currency with a precious metal to ensure acceptance in payment.

Historically, governments have sometimes maintained a reserve of gold or silver (or both) against their currency. It was thought that if the population could always return currency to the government to obtain precious metal instead, then the currency would be accepted because it would be thought to be 'as good as gold'. Sometimes the currency itself did contain precious metal, as in the case of gold coins.

For example, in the US, the treasury maintained gold reserves equal to 25 per cent of the value of the issued currency until the late 1960s, but American citizens were not allowed to redeem currency for gold; only foreign holders of US currency could do so. However, the US and most nations have long since abandoned this practice. Even with no gold backing, the US currency is still in high demand all over the world. This demonstrates that the view that currency needs precious metal backing is erroneous.

Legal tender laws

Another explanation offered for a currency's acceptance are legal tender laws. Historically, sovereign governments have enacted legislation requiring their currencies to be accepted in payments. Indeed, paper currency issued in the US proclaims 'this note is legal tender for all debts, public and private'; Canadian notes say 'this note is legal tender'; and Australian paper currency reads: 'This Australian note is legal tender throughout Australia and its territories.' By contrast, the paper currency of the UK simply says: 'I promise to pay the bearer on demand the sum of five pounds' (in the case of the £5 note). On the other hand, the euro paper currency makes no promises.

Further, throughout history there are many examples of governments that passed legal tender laws, but still could not create a demand for their currencies, which were not accepted in private payments, and were sometimes even rejected in payments by the government. In some cases, the penalty for refusing to accept a king's coin included the burning of a red hot coin into the forehead of the recalcitrant. Hence, there are currencies that readily circulate without any legal tender laws as well as currencies that were shunned even with legal tender laws. Further, as we know, the US dollar circulates in many countries in which it is not legal tender (and even in countries where its use is discouraged by the authorities).

Fiat currency

Modern currencies are often called **fiat currencies** because there is no promise made by government to redeem them for precious metal. Their value is proclaimed by 'fiat' (the government legislates a new issue of currency and announces that a coin is worth a half dollar without holding a reserve of precious metal equal in value to a half dollar). Many students on economics courses are shocked when they are first told that there is 'nothing' backing the currency in their pockets but the government's 'fiat'. While they had probably never contemplated actually taking the currency down to the treasury to exchange it for gold, they had found comfort in the erroneous belief that there was 'something' standing behind the currency, perhaps a reserve of precious metal that was available for redemption.

The UK currency's 'promise to pay the bearer on demand the sum of five pounds' appears to offer a sound basis, implying that the treasury holds something in reserve that it can use to make the promised payments. However, if one were to actually present to the UK government a five pound note, the treasury would simply offer another five pound note, or a combination of coins that sum to five pounds! Any citizen of the US or Australia would experience the same outcome at their own treasuries: a five dollar note can be exchanged for a different five dollar note, or for some combination of notes and coins to make five dollars. That is the extent of the government's 'promise to pay'!

If currency cannot be exchanged for precious metal, and if legal tender laws are neither necessary nor sufficient to ensure acceptance of a currency, and if the government's 'promise to pay' really amounts to nothing, then why would anyone accept a government's currency? Let us try to determine the answer.

Taxes drive the demand for money

One of the most important powers claimed by a sovereign government is the authority to levy and collect taxes (and other payments made to government, including fees and fines). Tax obligations are levied in the national money of account, for example, dollars in the US and Australia, yen in Japan, pounds in the UK and so on. Further, the sovereign government also determines what can be delivered to satisfy the tax obligation. In all modern nations, it is the government's own currency (usually in the form of its central bank reserves, as we explain next) that is accepted in payment of taxes.

Some taxpayers use cheques drawn on private banks to make tax payments, whilst others will transfer the funds electronically to the government. When government receives these cheques and transfers, it debits the reserves of the private banks, which are held at the central bank. Reserves are just a special form of government currency used by banks to make payments to one another and to the government. Like all currency, reserves are the government's IOU. Effectively, private banks intermediate between taxpayers and government, making tax payments in currency (reserves) on behalf of the taxpayers. Once the banks have made these payments, the taxpayer has fulfilled their obligation, so the tax liability is eliminated.[1]

We are now able to answer the question posed above: why would anyone accept a government's 'fiat' currency? The answer is that the government's currency is the main (and usually the only) thing accepted by government in payment of taxes and other obligations to the government. It is true of course that government currency can be used for other purposes: coins can be used to make purchases from vending machines; private debts can be settled by offering government paper currency; and government money can be hoarded in piggy banks for future spending. However, these other uses of currency are all subsidiary, deriving from government's willingness to accept its currency in tax payments. It is because anyone with tax obligations can use currency to eliminate these liabilities that government currency is in demand, and thus can be used in purchases or in payment of private obligations.

The government cannot easily force others to use its currency in private payments, or to hoard it in piggybanks, but government can force the use of its currency to meet the tax obligations that it imposes. For this reason, neither reserves of precious metals nor legal tender laws are necessary to ensure acceptance of the government's currency. All that is required is the imposition of a tax liability to be paid in the government's currency. The 'promise to pay' that is engraved on UK pound notes is superfluous and really quite misleading. We know that the UK treasury will not really pay anything (other than another note) when the five pound paper currency is presented. However, it will and must accept the note in payment of taxes. This is really how government currency is redeemed, not for gold but in payments made to the government. We will go through the accounting of tax payments in Chapter 20. It is sufficient for our purposes now to understand that the tax obligations to government are met by presenting the government's own IOUs to the tax collector.

We can conclude that **taxes drive money**. The government first creates a money of account (such as the dollar), and then imposes tax obligations in that national money of account. In all modern nations, this is sufficient to ensure that most debts, assets, and prices, will also be denominated in the national money of account. The government is then able to issue a currency that is also denominated in the same money of account, so long as it accepts that currency in tax payment. When we talk about the government 'issuing' currency, the most usual way in which this occurs is through government spending. We say **the government spends the currency into existence**. It can also make loans.

It is not necessary to 'back' the currency with precious metal, nor is it necessary to enforce legal tender laws that require acceptance of the national currency. For example, rather than engraving the statement 'This note is legal tender for all debts, public and private', all the US government needs to do is to promise 'This note will be accepted in the payment of taxes' in order to ensure its general acceptability within the US and even abroad.

In the Appendix to Chapter 2, we introduced the Buckaroos model which refers to the currency which US students acquire when they undertake hours of community service (CS) during each year of their degree programmes. Buckaroos (Bs) enable students to meet their tax obligations and this currency clearly has value, but is not backed by a previous metal. Buckaroos do not have widespread acceptability in the economy, because taxes are levied in $US by the government. However, it is quite conceivable that some transactions would occur between students in which Buckaroos are exchanged for dollars. Some students may be prepared to undertake additional hours of CS, whereas others may be prepared to buy Buckaroos with dollars, rather than undertaking the required CS work.

In Box 9.1, we illustrate the argument that fiat currencies have value, despite not being backed by precious metal, by reference to the use of paper currency in colonial Virginia in the late 18th century .

BOX 9.1 AN HISTORICAL NOTE: PAPER NOTES AND REDEMPTION TAXES IN COLONIAL AMERICA

The notion that taxes drive money can be demonstrated through examination of the history of coinage and of the issue of paper money. In his examination of colonial Virginia's use of paper currency Farley Grubb (2015) demonstrates the principle of imposing taxes for redemption of paper notes. The American colonies were prohibited by England from issuing coin, so as to protect the King's monopoly of coinage. The colonies obtained coin from exports, but as a major mercantilist power, England limited colonial exports to the raw materials they needed. The colonies had to import finished goods, shipping the coins back to England. The King also wanted to limit expenditures on his empire, so the colonies were largely responsible for funding their expenses, which included fighting wars with the French, the Canadians, and Native Americans. Colonial governments were hence chronically short of British coins, obtained through taxes such as poll taxes and taxes on exports of slaves and tobacco.

To increase fiscal capacity, the colonial governments began to issue paper money. Virginia's colonial government passed a series of acts to authorise the issue of treasury paper notes. Each law would include the total value of notes (denominated in Virginia pounds) to be issued and would set a date for final 'redemption' (the term used by Grubb as well as by the lawmakers). Interestingly, the law would also impose a new set of taxes at the time of the note issue:

> *Every paper money act included additional new taxes, typically a land tax and a poll tax, that were operative for a number of years. The number of years over which these new additional taxes were operative was chosen so as to generate enough funds to fully redeem the notes authorised by each respective paper money act. The date in each paper money act set for the final redemption of the notes authorised by that act closely matched the end to the taxing period set by that act. (Grubb, 2015: 27)*

The Paper Money Acts that allowed the treasury to issue notes also imposed new taxes, with the recognition that the purpose of the taxes was to 'redeem' the currency. In fact, colonial paper money could be 'redeemed' in two ways: payment of taxes or presentation at the treasury for payment in (British) coins. The treasury would spend the newly issued paper money into the economy and those receiving the treasury notes could use it to pay taxes, or spend it, or submit it to the treasury in exchange for coins.

What did the treasury do with the notes it received in tax payment? Grubb (2015: 17) reports that the "notes were removed and burned", not spent. This runs counter to the common belief that a government needs tax revenue in order to spend. The colonial case shows that government first had to spend before it collected tax revenue, and once it received the revenue, the government burned it rather than spent it.

Grubb shows that most taxes were paid using the paper money, and most paper money was 'redeemed' in tax payment:

> *A redemption tax of 10,327£VA was collected, of which 2,527£VA was in specie that was explicitly set aside in a dedicated account to be used to redeem notes brought to the treasury. The rest of the tax payments were burnt, implying that those tax payments were made in notes. Therefore, 76 percent of this tax was paid in notes, and 24 percent was paid in specie. (Grubb, 2015: 29)*

What about the notes that were not 'redeemed' by either method? They continued to circulate:

> *[A]t the final redemption date holders of the respective notes did not rush to the treasury to exchange them for specie. The notes continued in circulation and note holders could cash them in at the treasury at their leisure. Robert Nicholas Carter, Virginia treasurer after 1766, noted this behavior, 'Most of the Merchants as well as others, ... preferred them [Virginia's treasury notes] either to Gold or Silver, as being more convenient for transacting the internal Business of the Country' (William and Mary College Quarterly Historical Magazine, 1912: 235). (Grubb, 2015: 30)*

Likewise, Adam Smith (1937[1776]) had argued that if the colonies were careful to ensure they did not create too much paper money relative to taxes, it would not depreciate in value (indeed it might even circulate at a premium, he argued). Redemption of the notes in tax payment would remove them from circulation, keeping them scarce. Grubb argues that this was well recognised by the colonial government:

> *The Virginia legislature took note redemption and its effect on controlling the value of its paper money seriously. Such is illustrated in the March 1760 paper money act which stated, 'And whereas it is of the greatest importance to preserve the credit of the paper currency of this colony, and nothing can contribute more to that end than a due care to satisfy the publick that the paper bills of credit, or treasury-notes, are properly sunk, according to the true intent and meaning of the several acts of assembly passed for emitting the same; and the establishing a regular method for this purpose may prevent difficulties and confusion in settling the publick accounts... (Hening 1969, v. 7, p. 353)'. (Grubb, 2015: 27–28).*

This emphasises the fact that the notes were removed from circulation to protect the value of the government's paper currency, not to provide 'revenue' that government could spend. The problem with spending notes in excess of redemption would not be government insolvency, but rather inflation. The taxes were meant not to 'raise revenue' for spending. The government also realised it needed to receive a portion of tax revenue in the form of coin. This was to ensure that it could meet its promise to redeem notes for coin.

Redemption of the tax obligations by returning paper notes to the treasury not only 'redeemed' the colonial government (in the sense that its paper note debts were extinguished), but it also redeemed the taxpayers who owed taxes. The redemption is mutual and simultaneous: both the 'creditor' – the taxpayer – and the 'debtor' – the note-issuing treasury – were redeemed. At the same time, the 'debtor' taxpayer was redeemed of the duty to pay taxes to the 'creditor' treasury. The four entries on the balance sheets were all simultaneously wiped clean.

Creation of the notes preceded their redemption in tax payment. Note creation (through government spending) logically comes before note redemption (through taxation). Indeed, it would have been impossible for the colonists to pay the new taxes given the chronic shortage of coin unless the notes were issued and spent first. Nor would the government have needed to impose the new taxes if it was not going to spend the notes!

What this shows is that modern interpretations of 'redemption' are based on a narrow definition that applies when the issuer of a currency promises to 'redeem' that currency for either gold (gold standard) or a foreign currency (fixed exchange rate) at a promised exchange rate. Of course, there are issuers who make such a promise. However, the more common (and more fundamental) promise is that of accepting one's own liabilities in payments due, such as taxes owed to the issuer of a sovereign currency. Even in this case, the sovereign can also promise to 'redeem' the currency for gold or foreign currency (the Virginia colony promised redemption in English coin). We see this as an additional promise that applies in some cases, but a promise that is now rare among developed nations (the EMU nations are an exception[2]). The promise to accept its own currency in payment is the more common and indeed universal promise of 'redemption'. And it is sufficient to 'drive' a currency.

Financial stocks and flows are denominated in the national money of account

Financial stocks and financial flows are denominated in the national money of account. While working, the employee earns a flow of wages that are denominated in money, effectively accumulating a monetary claim on the employer (see Chapter 6). On payday, the employer eliminates the obligation by providing a wage payment via, say, an electronic transfer to the worker's bank account, that is a liability of the employer's bank. Again, that is denominated in the national money of account. If desired, the worker can draw on that bank deposit and receive the government's currency, again an IOU of the government.

Any disposable income that is not used for consumption purchases represents a flow of saving, accumulated as a stock of wealth. In this case, the saving is held as a bank deposit, that is, as financial wealth. These monetary stocks and flows are conceptually nothing more than accounting entries, measured in the money of account. We can easily imagine doing away with coins and paper notes as well as cheque books, with all payments made through electronic entries using computers connected via the internet. All financial wealth could similarly be accounted for without use of paper.

In Chapter 5, we carefully examined the definitions of stocks (for example, wealth) and flows (for example, income, spending, and saving), as well as the relationships between them.

The financial system as an electronic scoreboard

The modern financial system can be seen as an elaborate system of record keeping, a sort of financial scoring of the game of life in a capitalist economy. Financial scoring can be compared with a scoreboard at a sporting event. When a team scores, the official scorer awards points, and electronic pulses are sent to the appropriate combination of LEDs so that the scoreboard will show the number of points awarded. As the game progresses, points totals are adjusted for each team. The points have no real physical presence; they simply reflect a record of the performance of each team according to the rules of the game. They are not 'backed' by anything, although they are valuable because the team that accumulates the most points is deemed the 'winner', and perhaps rewarded with fame and fortune. Further, in accordance with applicable rules, points might be taken away after review by officials who determine that rules were broken and that penalties should be assessed. The points that are taken away don't really go anywhere; they simply disappear as the scorekeeper deducts them from the score.

Similarly, in the game of life, earned income leads to 'points' credited to the 'score' that is kept by financial institutions. Unlike a sporting contest, in the game of life, every 'point' that is awarded to one player is deducted from the 'score' of another, either reducing the payer's assets or increasing their liabilities. Accountants in the game of life are very careful to ensure that financial accounts always balance. The payment of wages leads to a debit of the employer's 'score' at the bank, and a credit to the employee's 'score', but at the same time, the wage payment eliminates the employer's implicit obligation to pay wages as well as the employee's legal claim to wages. So, while the game of life is a bit more complicated than the football game, the idea that record keeping in terms of money is a lot like record keeping in terms of points can help us to remember that money is not a 'thing' but rather is a unit of account in which we keep track of all the debits and credits, or 'points'.

When thinking about the 'scores' the currency-issuing government might record (via government spending the currency into existence), it doesn't make sense to say that the government can run out of money. That would be like saying a game must be terminated at some point before the scheduled end because the scorer had run out of scores to post on the scoreboard. We will come back to that point in later chapters.

9.3 Floating versus Fixed Exchange Rate Regimes

An **exchange rate** is the amount of currency A that can be purchased by a unit of currency B in what we call the foreign exchange market. We will consider these markets in more detail in Chapter 24. Government can allow its currency to be freely exchanged at whatever value the foreign exchange market determines (floating rates) or try to manage the exchange value, usually under multilateral agreements between nations (fixed rates). These different arrangements have implications for the conduct of economic policy, which we will briefly consider in this section.

In previous sections, we dealt with the case of governments that do not promise to convert their currencies on demand into precious metals or anything else. When a five dollar note is presented to the US Treasury, it can be used to pay taxes or it can be exchanged for some combination of notes and coins that sums to five dollars, but the US government will not convert it to anything else. Further, the US government does not promise to maintain the exchange rate of US dollars against other currencies at any particular level. This is the typical situation for most nations.

Most of this textbook will be concerned with sovereign currencies which operate with **floating exchange rates** against other currencies, so that they are not convertible at a fixed rate to another currency. Examples of such currencies include the US dollar, the Australian dollar, the Canadian dollar, the UK pound, the Japanese yen, the Turkish lira, the Mexican peso, the Argentinian peso, and so on.

What are the differences between fixed and floating exchange rates and what are the implications of this distinction?

The gold standard and fixed exchange rates

A century or so ago, many nations operated with a gold standard in which the country not only promised to redeem its currency for gold, but also promised to make this redemption at a fixed exchange rate. An example of a fixed exchange rate is a promise to convert 35 US dollars to one ounce of gold. For many years, this was indeed the official US exchange rate. Other nations also adopted fixed exchange rates, pegging the value of their currency either to gold, or after the Second World War, to the US dollar. For example, at the inception of the post-war system, known as the Bretton Woods system, the official exchange rate for the UK pound per US dollar was 0.2481 (on 27 December 1945). This is equivalent to a person receiving four US dollars for each UK pound presented for conversion. As all other currencies in the system were set relative to the US dollar, this also set their values relative to each other. So on 27 December 1945, 119.1 French francs exchanged for one US dollar, which meant that 480 francs were required to purchase one UK pound. In Chapter 24, we will learn how to interpret exchange rate quotations and to calculate various cross parities.

In order to make good on its promises to convert its currency at fixed exchange rates, each nation had to keep a reserve of foreign currencies (and/or gold). For example, if a lot of UK pounds were presented for conversion to US dollars (for example, by foreign central banks to the Bank of England), the UK's reserves of foreign (mostly dollar) currency could be depleted rapidly. There were three strategies that could be adopted by the UK government to avoid running out of foreign currency reserves, but none of them were very pleasant. They included: (a) alter the value of the pound against the US dollar, that is, devalue; (b) borrow foreign currency reserves; or (c) deflate the economy using higher interest rates and/or fiscal cutbacks to curtail imports and attract capital inflow.

Under this fixed exchange rate system, countries with trade deficits (exports less than imports) always had difficulties maintaining the agreed exchange parity because the trade deficit manifests in foreign exchange markets as an excess supply of the nation's currency relative to all other currencies. This is because when a nation sells exports, foreign buyers must supply their own currency in return for that nation's currency, and when a nation buys imports, it must supply its own currency in return for the currency of the nation from which it is importing. Thus, a trade deficit amounts to an excess supply of the deficit nation's currency in the foreign exchange market, which pushes the price (exchange rate) downwards. To arrest the decline in the exchange rate, the central bank is required to buy up its currency in the foreign exchange market using stocks of foreign currency, which eliminates the excess supply. However, nations with chronic trade deficits sooner or later ran out of stores of foreign currencies. These pressures eventually undermined the viability of the Bretton Woods system and it collapsed in August 1971.

Floating exchange rates

In August 1971, US President Nixon abandoned US participation in the fixed exchange rate system because the USA was unable to continue to guarantee conversion of US dollars into gold at the agreed price. Many countries followed suit. This meant that these governments no longer promised to convert their currency to another currency (or gold) at a fixed rate. As a result, the relative values of currencies against one another were allowed to float, meaning that they would be determined hour by hour according to forces of demand and supply in the foreign exchange market.

Today it is easy to convert most currencies, including floating currencies, into any other major currency at private banks and at kiosks in international airports. Currency exchanges enact these conversions at the current exchange rate set in international markets (minus the fees charged for the transactions). These exchange rates change day by day, or even minute by minute, fluctuating to match demand (from those trying to obtain the currency in question) and supply (from those offering that particular currency in exchange for other currencies).

The determination of exchange rates in a floating exchange rate system is exceedingly complex. The international value of the US dollar, for example, might be influenced by such factors as the demand for US assets, the US trade balance, US interest rates relative to those in the rest of the world, US inflation, and US growth relative to that in the rest of the world. So many factors are involved that no statistical model that can reliably predict movements of exchange rates has been developed yet.

What is important for our analysis however is that with a floating exchange rate, a government does not need to fear that it will run out of foreign currency reserves (or gold reserves) for the simple reason that it does not convert its domestic currency to foreign currency at a fixed exchange rate. Indeed, the government does not have to promise to make any conversions at all. In practice, governments operating with floating exchange rates hold foreign currency reserves, and they offer currency exchange services for the convenience of their financial institutions. However, the conversions are done at current market exchange rates, rather than keeping the exchange rate at a prescribed level.

Governments intervene into currency exchange markets to try to nudge the exchange rate in the desired direction. They also will use macroeconomic policy (including monetary and fiscal policy, as discussed in Chapter 20) in an attempt to affect exchange rates. Sometimes this works, and sometimes it does not. The point is that with a floating exchange rate attempts to influence exchange rates are discretionary. By contrast, with a fixed exchange rate government must use policy to try to keep the exchange rate fixed.

The floating exchange rate ensures that the government has greater freedom to pursue other policy goals, such as maintenance of full employment, sufficient economic growth, and price stability. How it might do that is discussed in later chapters.

9.4 IOUs Denominated in National Currency: Government and Non-Government

In previous sections we have noted that assets and liabilities are denominated in a money of account, which is chosen by a national government and given force through the mechanism of taxation. With a floating exchange rate, the government's own IOUs – its currency – are non-convertible in the sense that the government makes no promise to convert them to precious metal, to foreign currency, or to anything else. Instead, it promises to accept its own IOUs in payments made to itself (mostly tax payments, but also payments of fees and fines). This is the necessary and fundamental promise made: the issuer of an IOU must accept that IOU in payment. So long as the government agrees to accept its own IOUs in tax payments, the government's IOUs will be in demand (at least for tax payments, and probably for other uses as well).

Similarly, private issuers of IOUs also promise to accept their own liabilities. For example, if you have a loan with your bank, you can always pay the principal and interest on the loan by writing a cheque on your deposit account at the bank. Actually, all modern banking systems operate a cheque clearing facility so that each bank accepts cheques drawn on all other banks in the country. This allows anyone with a debt due to any bank in the country to present a cheque drawn on any other bank in the country for payment of the debt. The cheque clearing facility then operates to settle accounts among the banks. This topic will be discussed in detail in Chapter 20. The important point is that banks accept their own liabilities (cheques drawn on deposits) in payments on debts due to banks (the loans banks have made), just as governments accept their own liabilities (currency) in payments on debts due to government (tax liabilities).

Leveraging

There is one big difference between government and banks, however. Banks do promise to convert their liabilities to something. You can present a cheque to your bank for payment in currency, what is normally called 'cashing a cheque', or you can simply withdraw cash at the Automatic Teller Machine (ATM) from one of your bank accounts. In either case, the bank IOU is converted to a government IOU. Banks normally promise to make these conversions either 'on demand' (in the case of 'demand deposits', which are normal cheque accounts) or after a specified time period (in the case of 'time or term deposits', including savings accounts and certificates of deposit, known as CDs, perhaps with a penalty for early withdrawal).

Because banks make this promise to convert on demand, they must either hold reserves of currency, or have quick access to them. Their reserves take the form of vault cash plus deposits held at the central bank. Note that they need to hold only a small amount of reserves against their deposits because they know that redemptions (withdrawals) over any short period will be a tiny fraction of their total deposits.

The fraction of reserves against deposits is called the reserve ratio. We can think of deposits as **leveraging** the reserves. For example, in the USA, the ratio of reserves against bank deposits is around one per cent. This means the leverage ratio is 100 to one.

Banks hold a relatively small amount of currency in their vaults to handle conversions on demand, but most of their reserves take the form of deposits at the central bank. If they need more currency, they ask the central bank to send an armoured truck with the desired notes and coins. For our purposes here, bank reserves (deposits at the central bank) are equivalent to vault cash because a bank can immediately convert them to currency to meet cash withdrawals. There is no functional difference between cash held in bank vaults and reserve deposits held at the central bank. We can include both as currency, government liabilities with zero time to maturity.

Banks don't like to hold a lot of vault cash or reserves, nor do they need to do so in normal circumstances. Holding lots of cash on the premises could increase the attractiveness of a bank to robbers, but the main reason for minimising holdings is that it is costly to hold currency. The most obvious costs are the vault and the need to hire security guards. However, more important to banks is the fact that holding reserves does not earn much profit. Banks would rather hold loans as assets because debtors pay interest on these loans. For this reason, banks operate with high leverage ratios, holding a very tiny fraction of their assets in the form of reserves against their deposit liabilities. So long as only a small percentage of their depositors try to convert deposits to cash on any given day, this is not a problem. However, in the case of a bank run (in which a large number of customers try to convert their deposits to cash on the same day), the bank will have to obtain currency from the central bank. This can even lead to a lender of last resort action by the central bank in lending currency reserves to a bank facing a run. These are issues that we will address in Chapter 23.

Clearing accounts extinguish IOUs

There is another reason that banks hold reserves. When you write a cheque on your bank account to pay a bill, the recipient of the cheque will deposit it in their own bank, which is probably a different bank. Their bank will present the cheque to your bank for payment. This is called **clearing accounts**. Banks clear accounts using government IOUs, and for that reason banks maintain reserve deposits at the central bank. More importantly, they have access to more reserves should they ever need them, both through borrowing from other banks in the interbank market for reserves (an overnight market where banks lend to and borrow from each other), or through borrowing them from the central bank. All modern financial systems have developed procedures to ensure banks can get currency and reserves as necessary to clear accounts among themselves and with their depositors. The central bank is duty bound to provide banks with sufficient reserves should they fall short on any particular day.

When First National Bank receives a cheque drawn on Second National Bank, it asks the central bank to debit the reserves of Second National and to credit its own reserves. This is now handled electronically. Note that while Second National's assets will be reduced (by the amount of reserves debited), its liabilities (cheque deposit) will be reduced by the same amount. Similarly, when a depositor uses the ATM to withdraw currency, the bank's assets (cash reserves) are reduced, and its IOUs to the depositor (the liabilities in the deposit account) are reduced by the same amount.

Other business firms use bank liabilities for clearing their own accounts. For example, a retail firm typically receives products from wholesalers on the basis of a promise to pay after a specified time period (usually 30 days in the US). Wholesalers hold these IOUs until the end of the period, at which time the retailers pay by a cheque drawn on their bank account or by an electronic transfer from their account to the account of the wholesaler. At this point, the retailer's IOUs held by the wholesaler are cancelled.

Alternatively, the wholesaler might not be willing to wait until the end of the period for payment. In this case, the wholesaler can sell the retailer's IOUs at a discount (for less than the amount that the retailer promises to pay at the end of the period). The discount is effectively interest that the wholesaler is willing to pay to get the funds earlier than promised. The retailer effectively earns interest (the difference between the amount paid for the IOUs and the amount paid to the wholesaler to extinguish the IOUs). Again, the retailer's IOU is cancelled by delivering a bank liability (the holder of the retailer's IOU receives a credit to their own bank account). As we will see in Chapter 23, discounting is the basis of both commercial banking and of interest rates.

Pyramiding currency

Another important point is that private financial liabilities are not only denominated in the government's money of account, they also are ultimately convertible into the government's currency. As we have discussed, banks explicitly promise to convert their liabilities to currency (either immediately in the case of demand deposits, or with some delay in the case of time deposits). Other private firms mostly use bank liabilities to clear their own accounts. Essentially, this means they are promising to convert their liabilities to bank liabilities, 'paying by cheque' on a specified date (or according to other conditions specified in the contract). For this reason, they must have deposits, or have access to deposits, with banks to make the payments.

Things can get even more complex than this because there is a wide range of financial institutions (and even non-financial institutions offering financial services) that can provide payment services. These organisations can make payments for other firms, with net clearing among these 'non-bank financial institutions' occurring using the liabilities of banks. Banks in turn clear accounts using government liabilities. There could thus be 'six degrees of separation' (many layers of financial leveraging) between a creditor and a debtor involved in clearing accounts.

We can think of a pyramid of liabilities, with different layers corresponding to the degree of separation from the central bank. Perhaps the bottom layer consists of the IOUs of households that are held by other households, by firms engaged in production, by banks, and by other financial institutions. The important point is that households usually clear accounts by using liabilities issued by those higher in the debt pyramid, typically financial institutions.

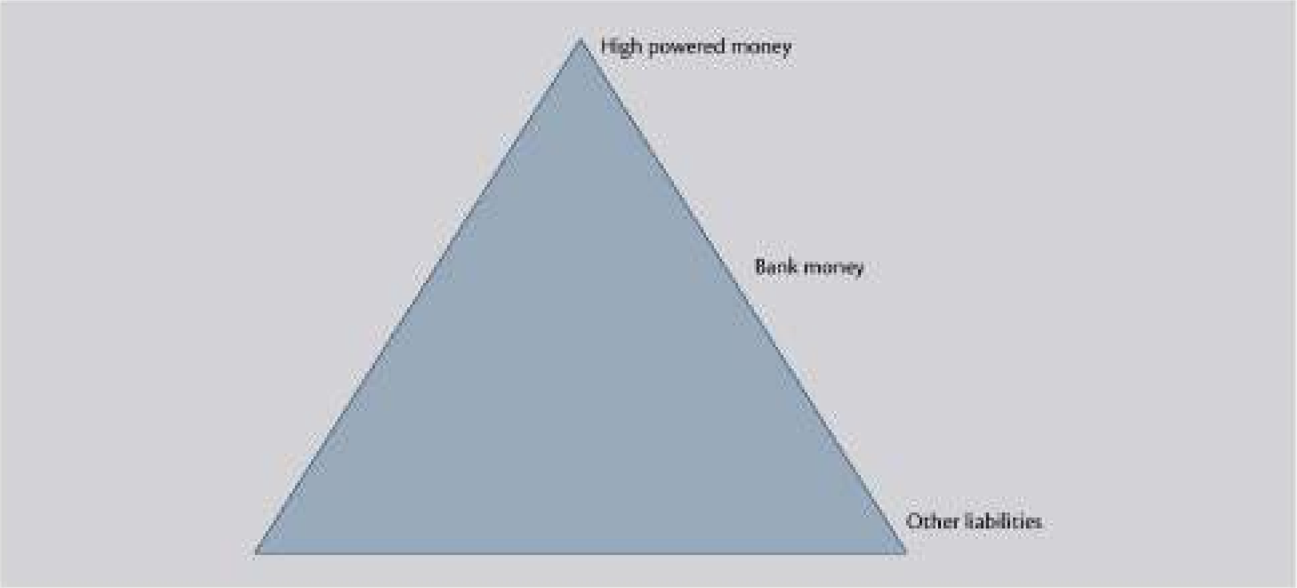
The next layer up from the bottom consists of the IOUs of firms engaged in production, with their liabilities held mostly by financial institutions higher in the debt pyramid (although some are directly held by households and by other production firms), and who mostly clear accounts using liabilities issued by the financial institutions, sometimes called shadow banks.

At the next layer we have non-bank financial institutions, which in turn clear accounts using the banks whose liabilities are higher in the pyramid. Banks use government liabilities for net clearing.

Finally, the government is at the top of the pyramid, with no liabilities higher than its non-convertible IOUs. The shape of the pyramid is instructive for two reasons. First, there is a hierarchical arrangement whereby liabilities issued by those higher in the pyramid are generally more acceptable. In some respects, this is due to higher creditworthiness because the government's liabilities are free of credit risk. As we move down the pyramid through bank liabilities, toward non-financial business liabilities, and finally to the IOUs of households, risk tends to rise, although this is not a firm and fast rule. Second, the liabilities at each level typically leverage the liabilities at the higher levels. In this sense, the whole pyramid is based on leveraging (a relatively smaller number of) government IOUs. This is a concept to which we will return in the next section.

Figure 9.1 shows a 'pyramid' (as per the concept developed by Hyman Minsky and Duncan Foley, and extended by Stephanie Bell) which provides a visual representation of the concept of leveraging. At the top of the pyramid are the government's liabilities, which we refer to as the monetary base, and constitute the sum of all bank

Figure 9.1 The Minsky–Foley pyramid



reserves held in the central bank clearing accounts and outstanding currency (notes and coins). At the bottom of the pyramid we include all other money-denominated liabilities (these could include the IOUs of non-financial firms as well as those of households).

9.5 Use of the Term 'Money': Confusion and Precision

Before concluding this chapter, we will briefly distinguish between our use of the term 'money' and the way it is often used. 'Money' is often used colloquially to refer to income, as in asking 'how much money do you make at your job?'. As was discussed in Chapter 5, income is a flow measured in nominal terms, that is, in the money of account. In this book, we will always carefully distinguish flows from stocks, and will not use the term 'money' in place of 'income'.

The term 'money' is also often used to indicate a particular liability, such as the demand deposit liability of a bank, or the currency IOU of the government. In fact, as we have discussed above, all financial liabilities are denominated in a money of account. It is thus rather arbitrary to call some of these 'money' and to exclude others. Further, each time one uses the term 'money' to refer to money-denominated liabilities in general, one must provide a list of those that are included as 'money' or a list of those that are excluded. Otherwise, we can never be sure what the speaker means.

Throughout this book, we will carefully distinguish between the money of account (the US dollar or the Australian dollar, for example), and specific money-denominated liabilities (demand deposits issued by banks or currency issued by the government, for example). The term 'money' simply refers to the unit of account chosen by government to denominate tax liabilities and payments made to government, the dollar in both the US and Australia.

As we have discussed, money does not have any physical existence but rather is the unit in which we can keep track of debts and credits, much as a 'point' is the unit of account used in a game of football to keep track of goals scored. Just as the score in football is denominated in goals, a coin is denominated in dollars (or fractions of a dollar). A goal in football takes a physical form (a player kicking the football into a specified area), but the points used to 'account' for the goal do not have any physical presence. In the same manner, a ten dollar note issued by a government has a physical form (a piece of paper imprinted with ink), but the ten dollars owed by the government that it 'accounts' for do not. We can think of the paper note as just the written record of the government's IOU. What does it owe you? The right to discharge your ten dollars of tax liabilities using the ten dollar 'record' of the government's IOU.

Conclusion

In this chapter we defined and examined the characteristics of a sovereign money system – one in which a government issues its own currency. We explained that most countries around the world today (and back through history) have each adopted their own currency because this is linked to a country's independence and fiscal sovereignty. We also explained why floating exchange rate regimes generally provide the greatest fiscal and monetary policy space. By contrast, pegging an exchange rate to a foreign currency or to gold generally reduces policy space and creates the possibility that a nation will be forced to default on the promise to convert its currency on demand. While many people believe that it is necessary to 'back' a currency with something of value (gold, foreign currency), this chapter introduced the concept that 'taxes (or other obligations) drive money'. In other words, if citizens need the government's currency to pay taxes, then that will be sufficient to guarantee that the currency will be accepted.

Finally, the concept of 'leveraging' of the state's currency was discussed. Private debts and credits are denominated in the government's money of account (the same money of account in which the currency is denominated). Some of these private debts (most notably, bank demand deposits) are made convertible on demand to the state's currency. These are 'cleared' using the state's currency. Other types of private debts are cleared using bank liabilities. This led to the notion of a 'pyramid' of liabilities with the government's liabilities at the top, 'leveraged' by those lower in the pyramid.

References

Grubb, F. (2015) "Colonial Virginia's Paper Money Regime, 1755–1774: A Forensic Accounting Reconstruction of the Data", Working paper No. 2015-11. Available at: http://lerner.udel.edu/sites/default/files/ECON/PDFs/RePEc/dlw/WorkingPapers/2015/UDWP2015-11.pdf, accessed 16 May 2017.

Smith, A. (1937[1776]) *The Wealth of Nations*, 5th edn, London: Methuen. Available at: http://www.econlib.org/library/Smith/smWNCover.html, accessed 4 April 2017.

Endnotes

1. The tax payment reduces the worker's financial wealth because their bank account is debited by the amount of the tax payment. At the same time, the government's asset (the tax liability owed by the worker) is eliminated when the taxes are paid, and the government's liability (the reserves held by private banks) is also eliminated. This is an example of the operation of the payments system, which will be analysed in greater detail in Chapter 20.
2. Each member nation issues Euro-denominated currency and bonds that are convertible to European Central Bank liabilities at par.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

10 MONEY AND BANKING

Chapter Outline

Learning Objectives

- Gain an understanding of definitions of the money supply and financial assets.
- Recognise the sharp distinction between the MMT and orthodox representations of the process of credit creation by banks.
- Be able to interpret a bank balance sheet and incorporate changes via flows of new transactions.

10.1 Introduction

In this chapter, we have several objectives. First we will introduce students to commonly used definitions of the **money supply**. Frequent reference has been made in earlier chapters to the purchase or sale of financial assets by both the government through the central bank and treasury, as well as by banks. Here we will provide students with a clear understanding of the generic characteristics of these **financial assets**. We then devote space to developing an understanding as to how banks behave in a modern monetary economy. In the process, we will expose some long-standing myths about the role that banks play in the operation of the financial system.

10.2 Some Definitions

Monetary aggregates

Economists and commentators draw inferences about the economy from trends over time in monetary aggregates. Several measures of monetary aggregates have been devised over the years, but there is some variation across countries in what components are included under each measure. The different measures published by central banks are sometimes summarised as *M0*, *M1*, *M2*, *M3* and *M4* and reflect varying degrees of liquidity (convertibility into cash). It is common to consider the highest-liquidity measures M0 and M1 as **narrow money** while M2, M3 and M4 are considered to measure **broad money**.

M0 is also termed the **monetary base**. In countries such as Australia and the UK it includes circulating notes and coins held by the non-government sector, including banks; the deposits of banks with the central bank (generally called 'reserves'); and other central bank liabilities to the non-government sector. In the USA the monetary base is defined in the same way, but the term M0 is not used. The monetary base is the most liquid measure of the money supply and is also sometimes referred to as **high powered money** (HPM), due to its use as a reserve that is leveraged by private banks that issue their own money-denominated liabilities such as deposits.

M1 typically comprises notes and coins in circulation plus current bank deposits held by the private non-bank sector. In some nations, it includes travellers' cheques and deposit accounts that cheques can be written against. It is also a liquid measure of the money supply because its components are readily available to be used for spending on goods and services.

The US Federal Reserve defines M2 as M1 plus most savings accounts, money market accounts, retail money market mutual funds, and small denomination time deposits (certificates of deposit of under $100,000). M2 is a less liquid measure of the money supply and movements in it are typically used to forecast inflation.

M3 broadens the narrow measures to include less liquid components such as long-term time deposits. Even broader still are M4 measures which add other illiquid assets to the aggregate, such as borrowings from the private sector by non-bank financial intermediaries.

Not all measures are published by all central banks. The US, for example, only publishes the monetary base, M1 and M2. In the UK, there are only two official money supply measures (M0 and M4). The European Central Bank publishes M1, M2 and M3, while in Australia the central bank publishes M0, M1, M3 and M4 (or broad money measure).

10.3 Financial Assets

If a household engages in saving (a flow per period of time) over a number of months or years, then it will accumulate a growing stock of wealth. The household needs to decide whether to continue to add its saving to its existing deposits at its bank or put together a portfolio of financial assets which have different degrees of risk, for example, stocks (shares) or bonds, and are also denominated in the money of account.

Treasuries in modern economies issue bonds which are debt of various maturities; also called securities. These financial assets are bought and sold by the central bank, private banks and the private sector. Private entities (for example, corporations) also issue bonds.

In general, the bond acknowledges that the **issuer** is indebted to the **bondholder**. The bond issuer must pay interest to the bondholder on a periodic basis, and repay the principal (face value of the bond) when the bond matures. Bonds represent wealth for bondholders.

Thus, a **bond** is a **formal contract** to repay a loan (IOU) with interest at fixed intervals. The bondholder is the lender (creditor). The borrower (debtor) issues the bond and the **coupon** is the interest rate paid on the **face value** of the bond, and usually printed on the bond. In these cases, the periodic interest payments are constant.

The **issue price** is what investors pay for the bond when it is first issued. Later, bonds may be traded; at a premium (above par, if good quality, so that there is minimal default risk by the issuer) or at a discount (price below par). The bonds of a currency-issuing government carry zero default risk because such a government can always meet its outstanding liabilities. For this reason, these bonds are very desirable in times of uncertainty.

A **consol** is a special type of bond called a perpetuity, which means there is no maturity date. Interest is paid on this asset forever.

When we talk of the **government bond market** we need to differentiate between the primary and secondary bond markets. A **primary market** is the institutional machinery by which the government sells debt to the non-government sector. While many mistakenly believe that the issuance of bonds in the primary market is designed to raise funds to facilitate government spending, the reality is that currency-issuing governments are not financially constrained (see, for example, Chapter 1) and therefore we must seek a different explanation of why such governments issue debt at all to the non-government sector. We deal with those questions in more depth in Part E of this textbook (*Economic Policy in an Open Economy*).

A **secondary market** is where existing government bonds are bought and sold by interested parties after the bonds enter the monetary system via the primary market. The same arrangements apply, for example, to private share issues (also called equities or stocks). The company raises funds via the primary issuance process and then its shares are traded in secondary markets.

Government bonds are thus negotiable because ownership of the certificate can be transferred (sold) to another owner in the secondary bond market. However, it is important to understand that once the bond is issued, subsequent trading has no impact at all on the volume of financial assets in the system since it just shuffles this wealth between wealth holders.

The process of issuance in primary markets varies across nations and has also varied over time. A typical arrangement in the past was that government bonds would be sold to selected dealers (for example, banks) on a periodic basis in the primary market. Government would determine how much debt it wanted to issue (expressed in the money unit of account) and set a yield it was prepared to pay to the purchasers. The terms offered by this 'take it or leave it' approach might not be attractive to the non-government sector at the time of offer so any shortfalls in purchases of what the government desired to issue would be taken up by central banks. This is a case of government issuing debt to itself, which brings into question the whole logic of issuing debt.

In the late 1970s, the dominant school of economic thought was Monetarism, which erroneously claimed that central bank purchases of debt would be inflationary. Governments fell prey to that logic and started to devise ways to preclude their central banks from purchasing unsold debt. Governments would thus set yields and sell as much debt as possible, but would continuously adjust the yields up or down to meet the market requirements and ensure that there were no discrepancies between net spending (fiscal deficits) and bond sale revenue.

This system gave way to a purer auction system which avoided any claims that the government was manipulating yields, again in response to calls for more 'free' market activity. These auction or tender systems became dominant internationally. In general terms, the treasury would announce the terms of the auction, including how much debt was available for sale, the maturity dates of the debt, and the coupon rate (the periodic interest to be paid on the face value of the bond). The issue would then be put out for tender and then the bond dealers in the primary market would determine the final price of the bonds issued – thus taking discretion away from the elected government in terms of the yields that it would pay on government debt issuance.

As an example, imagine a $1,000 bond had a coupon of five per cent, meaning that you would get $50 per annum until the bond matured, at which time you would get $1,000 back. At the time of issue, the bond market dealers desired a yield of six per cent to satisfy their profit expectations. In this case, the initial specification is unattractive. Prior to the adoption of an auction system, private bond dealers would avoid purchasing the bond under such conditions. But under the auction system they could put in a purchase bid lower than the $1,000 to ensure they got the six per cent return sought on the price that they were willing to pay.

It is important to understand that there is an **inverse relationship between the traded price of a fixed income bond and its yield (rate of interest)**. Why is that so? The general rule for fixed income bonds is that when their prices rise in secondary markets, the yield falls and vice versa. This is because if one pays more to purchase a bond, the coupon payments represent a lower return on the purchase price; on the other hand if one pays less, then the coupon payments represent a higher return. Furthermore, the price of a bond can change in the marketplace according to interest rate fluctuations, even though the bondholder will still only get the face value of the bond back upon maturity.

When interest rates rise elsewhere in the economy, the price of previously issued bonds falls because they are less attractive in comparison to the newly issued bonds, which are offering higher coupon rates (reflecting current interest rates). When interest rates fall, the price of older bonds increases as they become more attractive given that newly issued bonds offer a lower coupon rate than the older higher coupon rated bonds.

The government department that manages these auction processes receives tenders from the bond market traders in the primary bond market. These will be ranked in terms of price (and implied yields desired) and the quantity requested in dollar terms. The bonds are then issued in order of the highest price bid down until the volume the government desires to sell is achieved. So, the first bidder with the highest price (lowest yield) gets their desired volume (as long as it doesn't exhaust the whole tender, which is unlikely). Then the second bidder (higher yield) receives their allocation and so on. In this way, if demand for the tender is low, the final yields will be higher and vice versa.

Bonds are also issued and sold in primary markets by state or provincial governments, multinational and local companies, credit institutions and other public bodies. Companies can finance new capital investment by one or more of the following: (i) issuing bonds, (ii) using retained profits, and (iii) launching a new share issue.

Treasuries and other institutions issue bonds with different times to maturity. For example, the US Department of the Treasury issues bonds of one month, three month, six month, one year, two year, three year, five year, seven year, ten year, twenty year and thirty year duration; a ten year treasury bond matures in ten years, and so on.

Yield concepts in fixed income investments

The yield indicates the return that will be returned from the investment and is usually expressed in percentage terms. There are several concepts of yield that are used in the markets.

- **Coupon or Nominal Yield** – If a bond has a face value of $1,000 and is paying eight per cent in interest (the coupon rate), then the nominal yield is eight per cent. The investor will thus receive $80 per annum until maturity. The coupon yield remains constant throughout the life of the bond.
- **Current Yield** – Suppose you purchase an eight per cent $1,000 bond for $800 in the secondary market. Irrespective of the price you pay, the bond entitles you to receive $80 per year in coupon payments. But unlike the previous example, the $80 payment per year until maturity represents a higher current yield than eight per cent on your investment because it is based on your purchase price of $800. The actual yield is $80/$800 = ten per cent. So, to compute current yield you simply divide the coupon by the price you paid for the bond. In general, if you buy the bond at a discount to face value, the current yield will be greater than the coupon yield, and if you buy at a premium then the current yield will be below the coupon yield.
- **Yield to Maturity (YTM)** – The current yield does not consider the difference between the purchase price of the bond and the principal payment at maturity. YTM considers that in addition to earning interest, an investor can make a realised capital gain or loss by holding the bond until its maturity date. YTM is a measure of the investor's true gain over the life of the bond and is the most accurate method of comparing bonds with different maturity dates and coupon values.

BOX 10.1 WORKED YIELD EXAMPLE

Assume you pay $800 for a $1,000 face value bond in the secondary market. The $200 discount on the face value is considered income or yield and must be included in the yield calculations. Assume that the eight per cent $1,000 bond has five years left to maturity when bought for $800.

A comparison of three yield concepts gives:

- Coupon yield of eight per cent ($80 income flow divided by $1,000 face value).
- Current yield of ten per cent ($80 income flow divided by $800 discounted purchase price).
- YTM of 13.3 per cent. The working is given in the main text below.

The computation of YTM is complex and can be simplified to the following rule of thumb:

$$YTM = (C + PD)/[0.5 \times (FV + P)]$$

where C is the coupon, PD is the prorated discount, FV is the face value, and P is the purchase price. PD is the difference between the face value and the purchase price, divided by the number of years to maturity. If the bond is trading at a premium, the PD is negative which means that the YTM is less than the coupon yield.

Using the data in the Worked Example, therefore:

$$YTM = [80 + (200/5)]/[0.5 \times (\$1{,}000 + \$800)] = \$120/\$900 = 13.3 \textit{ per cent.}$$

When bond traders talk about yield they are usually referring to the YTM measure which is the only measure that takes into account the effect of principal price, coupon rate, and time to maturity of a bond's actual yield.

There are two ways we can use data on yields for government bonds of different maturities to assess the state of the economy and the degree to which the non-government sector expects inflation to increase in the future. We have seen that rising yields signal weakening prices due to falling private demand for the assets in question. This could reflect a strengthening economy with investors being prepared to acquire more risky assets and less very safe ones. This is also usually when the central bank pushes up the target interbank rate and bond yields more or less follow (see Chapter 20). Further, we can use the movements in yields to gauge what is happening to inflationary expectations in the non-government sector. Rising yields on longer-term maturities indicate that the private markets sense inflation will rise in the future and so they desire to protect real yields by increasing the nominal yields on the bonds.

The second way of looking at the yields is to consider the **yield curve**. The yield curve is a graphical depiction of the **term structure of risk-free interest rates** and plots the maturities of different government bond on the horizontal axis against their respective yields (rates of return) on the vertical axis (see Figure 10.1, which shows the US Treasury yield curve for 3 February 2016).

There are various theories about the yield curve and its dynamics. All share some common notions, in particular that the higher is expected inflation, the steeper the yield curve will be, other factors being equal.

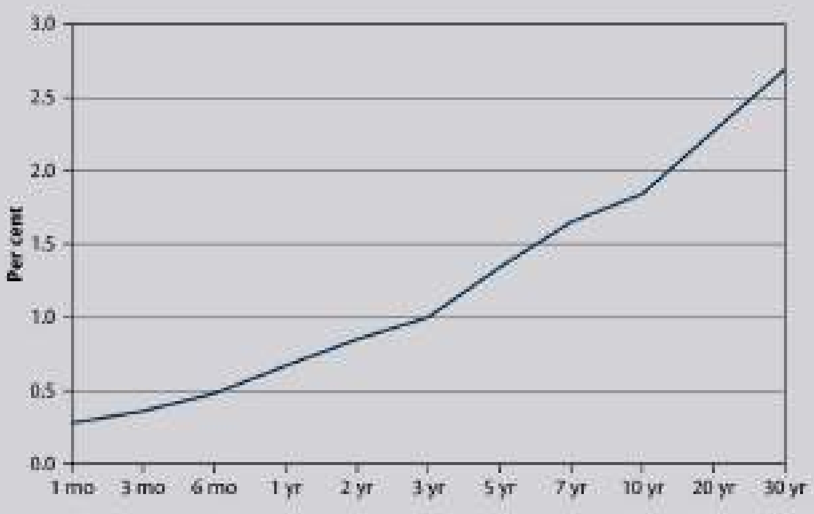
The basic principle linking the shape of the yield curve to the economy's prospects is explained as follows. The short end of the yield curve reflects the interest rate set by the central bank, which sets the competitive rate for cash (highly liquid assets) in the economy. As the short-term interest rate rises (falls), the yields on other less liquid assets will follow suit. The steepness of the yield curve then depends on the yield of the longer-term bonds, which are set by the market. But the short end of the curve is the primary determinant of its slope. In other words, the curve steepens mainly because the central bank is lowering the official cash rate, and it flattens mainly because the central bank is raising the official cash rate.

Bond traders link the dynamics of the yield curve to their expectations of the future economic prospects that are expected to influence central bank interest rate policy. It must be remembered that if central banks raise interest rates, then this will tend to cause prices of bonds to fall. This is called a **capital loss**. The prices of bonds with the greatest term to maturity will tend to be affected the most, so longer-term bonds are generally subject to the greatest risk of capital loss. For this reason, there may be a link between inflation expectations, expectations of central bank policy, and prices and yields on longer-term bonds.[1]

In summary, there are three shapes that the yield curve can take:

- **Normal** – Under normal circumstances, short-term bond rates are lower than long-term rates. The central bank attempts to keep short rates down to keep levels of activity as high as possible and bond investors desire premiums in longer-term maturities to protect them against capital losses. Thus, the yield curve is upward sloping, as in the case shown for the US in Figure 10.1.
- **Inverted** – Sometimes, short-term rates are higher than long-term rates and the yield curve is said to be inverted. When the economy starts to overheat, the expectations of rising inflation that might induce the central bank to raise its target interest rate lead to higher bond yields being demanded on assets with longer-term

Figure 10.1 US Treasury yield curve (3 February 2016)



Source: Authors' own. Data from Daily Treasury Yield Curve Rates, Resource Center, US Department of the Treasury (https://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/default.aspx).

maturities. The central bank might respond to the building inflationary pressures by raising short-term interest rates sharply. Although bond yields might rise, the significant tightening of monetary policy causes short-term interest rates to rise faster, resulting in an inversion of the yield curve. The higher interest rates may then lead to slower economic growth.

- **Flat** – A flat yield curve is seen most frequently in the transition from positive to inverted and vice versa. As the yield curve flattens, the yield spreads drop considerably. A yield spread is the difference between say, the yield on a one year and a ten year bond. What does this signal about the future performance of the economy? A flat yield curve can reflect a tightening monetary policy (short-term rates rise). Alternatively, it might depict a monetary easing after a recession (easing short-term rates) so the inverted yield curve will flatten out.

Movements in the yield curve are thus closely watched by economists due to the information that they convey about the general health of the economy, possible central bank interest rate adjustments, and inflationary expectations in the non-government sector.

BOX 10.2 THE ORTHODOX APPROACH TO NOMINAL INTEREST RATE DETERMINATION: THE FISHER EFFECT

One risk in holding a fixed coupon bond with a fixed redemption value is purchasing power risk.

Orthodox economists who adopt the loanable funds approach to interest rates believe that most people would prefer to consume now rather than later. To encourage forgone consumption now, a yield on savings must be provided by markets. The yield is intended to allow a person to consume more in the future than has been sacrificed now. But if the prices of goods and services increase in the meantime, then inflation could completely wipe out any gain in real consumption, so that the real interest rate is zero.

Consider a person who invests in a one year $1,000 coupon treasury bond with an expected single coupon payment of $100. The individual will expect to get $1,100 on the redemption date.

Assume that over the holding period, prices rise by ten per cent. At the end of the year, a basket of goods that previously cost $1,000 would now cost $1,100. In other words, the investor is no better off at the end of the year as a result of the investment. The nominal yield has been offset by the price inflation. Orthodox economists believe that investors are motivated by 'real' returns, not by nominal returns. This is because they view the decision to invest as coming from consumers who choose whether to consume now or to consume in the future, with consumption taking the form of real goods and services. If such savers do not take account of inflation, their future real consumption will be less than they desired.

Orthodox economists propose that the nominal interest rate must equal a real interest rate plus expected inflation. The real interest rate is supposed to be the market-determined real return that will equate the saving supply of funds with the investment demand for funds. It is thus a real equilibrium interest rate. However, as contracts are written in nominal terms, that is, in terms of nominal interest rates, the nominal rate must include compensation for the expected inflation rate. This addition to the real rate as inflation expectations rise is called the Fisher effect, named after the American economist Irving Fisher, who identified this relationship in the 1930s. Many market participants believe this applies to bond markets, and there is a strong belief that nominal yields are adjusted by markets to preserve purchasing power.

Purchasing power risk increases as the maturity lengthens. This is one reason many economists believe that longer maturity rates will generally be higher. The market yield is equal to the real rate of return required plus compensation for the expected rate of inflation. If the inflation rate is expected to rise, then market rates will rise to compensate. In this case, we would expect the yield curve to steepen, given that the Fisher effect will impact more significantly on longer maturity bonds than at the short end of the yield curve.

10.4 What Do Banks Do?

The neoclassical view: the money multiplier

In most textbooks, banks are presented as financial intermediaries that take in deposits, hold a small fraction of these in the form of reserves, and lend out the remainder. The causality is from deposits to reserves to loans. If each bank follows these principles in making loans, aggregate lending expands through the 'deposit or money multiplier'. For the moment assume that all banks are required to maintain a ratio of reserves to deposits of ten per cent. This might be to enable them to readily respond to a loss of reserves resulting from spending by customers (on goods and services, say) whose sellers bank elsewhere, or customers seeking to hold additional cash.

This is how the neoclassical school of thought describes the operation of the **money multiplier**:

i) Assume that a customer deposits \$100 in Bank A;
ii) Bank A retains \$10 of reserves to conform to the required reserves-to-deposit ratio of ten per cent. To expand its loan portfolio and increase profits, the remaining \$90 is loaned to a customer whose deposits rise by \$90;
iii) The customer spends these deposits and the recipient of the funds (seller) deposits \$90 in Bank B;
iv) Bank B then lends $0.9 \times \$90 = \$81$ (keeping ten per cent, that is, \$9 as additional reserves, as required) to a customer to finance their expenditure and so on.

At each stage the amount lent and then spent diminishes. It can be readily shown that if this was the way the banking system operated, then \$900 of additional loans are created. With the initial new deposit, this means that deposits have risen by a total of \$1,000 and are 'backed' by \$100 of reserves, thereby conforming to the required ten per cent ratio.

This example is what the mainstream textbooks call a **fractional reserve banking system**. It purports to explain how banks create money, which increases the M1 money supply due to the increase in current deposits. In terms of the initial deposit of \$100, the multiplier is 10, which is the inverse of the required reserves-to-deposit ratio of 0.1. A smaller money multiplier results if the non-government sector chooses to hold more cash when credit is created.

Note that no individual bank 'creates money' in this example, but the system as a whole 'multiplies' the initial deposit of $100 into $1,000. At each step, each bank simply lends out 90 per cent of the deposit it has received, keeping ten per cent as reserves. According to mainstream textbooks, the 'magic' results from fractional reserve banking. The larger the fraction of a deposit that must be retained as reserves, the smaller the multiplier effect. Following this logic, if the reserve ratio were zero (no reserves held against deposits), the banks would create an infinite amount of deposits after the deposit of just one dollar.

The standard textbook example is typically assumed as a ten per cent ratio, so that students can readily calculate a money multiplier equal to ten! On 12 April 1992, the US Federal Reserve Bank, for the first time in history, set the required reserve ratio on demand deposits at the magical ten per cent, making theory appear to coincide with reality. But that coincidence did not make the theory correct. As we will see next, the money multiplier as a description of modern banking is a myth, and bears no relation to how banks operate in the real world.

To summarise the dominant neoclassical view, banks are conceived as being financial intermediaries that maximise profits. They take in deposits to build up reserves so that they can then on lend the deposits at a higher interest rate. However prudential regulations require that they maintain a minimum reserve-to-deposit ratio. The fractional reserve requirements mean that the resulting credit creation process is finite.

In addition, many economists still believe that the monetary base, which consists of bank reserves and cash held by the non-government sector, is under the control of the central bank. Thus, by controlling the size of the monetary base and setting the required reserve ratio, the central bank is considered to be able to control the size of the money supply or the quantity of money.

Thus in the neoclassical narrative, the money supply is considered **exogenous** and determined by the central bank. This is an important claim because it has underpinned arguments that central banks can cause inflation by allowing the money supply to grow too quickly. From this follows the Quantity Theory of Money's (QTM) policy recommendation that the central bank can fight inflation by slowing money growth. As we will see in Chapter 20 (and analyse in Chapter 23), the QTM is a flawed conception of the inflation generation process. We will also demonstrate that the central bank does not have the capacity to control the money supply in a normally functioning money system.

The implication of the operation of the money multiplier is that a bank would forgo profitable loan opportunities if it did not have sufficient reserves to enable additional credit creation. Some allowance is made for discretion: the deposit multiplier is claimed to be a function of interest rates and interest rate differentials, bank preferences regarding their holdings of excess reserves, and also public preferences regarding their holdings of cash, as noted, and time deposit and demand deposit ratios. However, as Brunner (1968) 'demonstrated', these factors are of only minor importance.

MMT representation of the credit creation process

The neoclassical characterisation of the credit creation process, which is driven by fractional reserve requirements, is not an accurate depiction of the way banks operate in a modern monetary economy characterised by a fiat currency and a flexible exchange rate.

In the real world, the business of banking is complicated but is, in some respects, similar to that of other profit-seeking firms. Like these other firms, banks seek to earn profits and thereby generate returns for shareholders. Making loans secures profits, as long as the banks are paying a lower rate of interest on the funds that they borrow than they receive from their customers who take out loans.

First, a necessary condition for credit creation is that there are non-bank firms and/or households who are seeking loans to finance their planned spending on goods, services or assets. Second, some of these entities must be considered creditworthy by the banks, so that there is a high probability that the loan will be repaid in full. What constitutes creditworthiness varies over the business cycle and lending standards tend to become more lax in boom times as banks chase market share. Third, the banks must anticipate that there is profit to be made by making these loans, as described above.

Banks make loans independently of their reserve positions (that is, their holdings of reserves, relative to their liabilities). After originating loans they will borrow additional reserves if required by law or for clearing purposes.

Bank managers generally neither know, nor care, about the aggregate level of reserves in the banking system. Certainly, no loan officer ever checks the individual bank's reserve position before approving a loan. Bank lending decisions are affected by the price of reserves and expected returns, not by reserve positions. If the spread between the rate of return on an asset (a security or a loan) and the cost of borrowing reserves is wide enough, even a bank that is already deficient in reserves will purchase the asset or make a loan and cover the reserves needed by purchasing (borrowing) reserves in the **interbank market**. The interbank market connects the banks which lend reserves to and borrow reserves from each other when needed.

The important point is that when a bank originates a loan to a firm or a household, it is not lending reserves. Bank lending is not easier if there are more reserves, just as it is not harder if there are less. Bank reserves do not fund money creation in the way that is claimed in the money multiplier and fractional reserve deposit story; banks do not wait for deposits to come in before they make loans.

The main difference between banks and other types of firms involves the nature of the liabilities. Banks 'make loans' by purchasing the IOUs of 'borrowers'. This results in a bank liability, usually a demand deposit, at least initially, that shows up as an asset of the borrower. Thus, a customer of a bank who secures a loan is simultaneously a 'creditor' of the bank, due to holding a demand deposit, but also a 'debtor' to the bank. These creditors will almost immediately exercise their right to use the newly created demand deposits as a medium of exchange for purchases of goods and services, or assets. Bank liabilities (bank deposits) are used by households and non-bank firms for transactions in the form of cheques or transfers. Customers can also redeem demand deposits at par (dollar for dollar) against fiat money (which is guaranteed by the government) to enable cash to be used for purchases or making payments that are due. The government will also accept some kinds of bank liabilities in payment of taxes.

In turn, bank reserves are used for payment (or interbank settlement) among banks and for payments made to the central bank. Thus, when bank 'creditors' draw down their demand deposits, by either spending or choosing to hold more cash, a corresponding loss of reserves for the individual bank results. The bank may then either sell an asset or increase its liabilities (borrowing additional reserves) to cover the loss of reserves.

The interbank market (called the federal funds market in the US) functions to shuffle the reserve balances that the member (private) banks keep with the central bank to ensure that each of these banks can meet its reserve targets, which might be simply zero balances at the end of a specified period of time (that for simplicity, we could assume is a day). In aggregate, however, such activities only shift reserves from one bank to another. If more reserves are needed in total, they must be supplied by the central bank.

Far from waiting for deposits before they create loans, banks in the real world expand their balance sheets by lending as described below.

Loans create deposits

Loans create deposits that are then backed by reserves after the fact. The process of extending loans (credit), which creates new bank liabilities, is unrelated to the reserve position of the bank. In the pursuit of profit, banks take applications from creditworthy customers who seek loans and assess them according to the verity of the application, although in the lead-up to the Global Financial Crisis of 2008, the validation process became very lax.

The only thing that constrains the bank loan desks from expanding credit is a lack of creditworthy applicants, which can be due to banks raising the qualifying standards in times of pessimism, or can occur if creditworthy customers are loath to seek loans because of future uncertainty.

The major insight is that any balance sheet expansion that leaves a bank short of the required reserves may affect the return it can expect on the loan. This is a consequence of the 'penalty' rate the central bank might exact through the discount window (the central bank facility for lending to banks in need of reserves) should the bank fall short of the reserves it requires at the end of the day to cover the claims on it. However, it will never impede the bank's capacity to make the loan in the first place. It is thus quite wrong to assume that the central bank can influence the capacity of banks to expand credit by adding more reserves into the system. We will address this proposition in more detail in Chapter 23.

Banks do not loan out reserves

A corollary of the 'loans create deposits' insight is that banks do not loan out reserves, which raises the question of what role do bank reserves actually play?

Banks must hold reserve balances with the central bank as part of the **payments system**. The reserves are used to make interbank payments. Each day millions of transactions are reconciled (settled) through these interbank payments. For example, cheques drawn on Bank A and deposited at Bank B will see funds transferred from Bank A's reserve balances to those of Bank B.

If a particular bank finds itself short of the quantity of reserves necessary to resolve all the daily claims against it, then it can first try to borrow reserves from other banks that might have excess reserves in relation to their requirements on that particular day. But, as we will see in Chapters 20 and 23, an overall shortage (or excess) of reserves across the banking system must be rectified by the central bank which provides reserves to banks (in the case of a shortage) or may drain them from the system in the case of an excess. This central bank intervention is what we refer to as its liquidity management role and allows the bank to manage the overall level of reserves so that it is consistent with its interest rate target. For example, if on any particular day there is an excess of reserves (over and above the quantity required to settle transactions) and the central bank does not offer any competitive return on them, banks holding those excesses will try to loan them out overnight, which has the effect of driving down the short-term interest rate. The central bank must drain those reserves (by selling government bonds to the banks in return for debits to the reserve accounts) to ensure the overnight interbank interest rate remains equal to its desired policy (target) rate. We will learn more about this in Chapter 23.

Endogenous money

We have stated that unlike the story presented in neoclassical textbooks, in the real world the central bank cannot control the money supply. In other words, the money supply is **endogenous money** in the sense that the supply of bank money is determined 'endogenously' by the demand for bank loans, plus the willingness of banks to lend (which gives rise to the creation of deposits). The neoclassical theory erroneously believes that the money supply is exogenous and determined through the money multiplier interacting with the monetary base, which neoclassicists believe to be under the control of the central bank.

The demand for bank loans is determined by the spending decisions of private economic agents (including decisions regarding asset purchases). These can be affected, but only very indirectly, by the loan rate of interest. Banks supply loans only because someone is willing to 'borrow' bank money by issuing an IOU to banks. This means that the interest rate cannot be determined by the supply of and demand for loans since supply and demand are not independent. Rather, banks are price setters in short-term retail loan markets. They then meet the demand for loans with some quantity rationing at that price. In other words, some requests for loans are refused, even where aspiring borrowers claim to be willing and able to pay the going interest rate.

There can be several reasons for such quantity rationing of large segments of the population. Banks might worry about the default risk of some borrowers but might not be able to raise interest rates sufficiently to cover the default risk. Quantity rationing is then superior to price rationing, that is, raising the interest rate charged to some borrowers. Also, banks probably have better information than do borrowers about default risks. For example, the borrower who wishes to open a new restaurant might not have accurate information about bankruptcy rates in the industry or might simply be overly optimistic. On the other hand, banks can never know the future, so must operate based on rules of thumb (for example, informal rules that restrict loan size). Some quantity rationing can even be irrational, perhaps discriminatory, because banks have traditionally forgone certain kinds of loans or are reluctant to lend to certain groups in the community. The key point is that the supply of loans does not simply adjust to the demand for loans at some interest rate.

The short-term retail interest rates can be taken as a mark-up over short-term wholesale interest rates. Exactly what determines the mark-up (and whether it is variable) is controversial, but not important to our analysis here (see Moore, 1988).

Wholesale interest rates, finally, are under the influence of central bank policy. Individual banks use wholesale markets to rectify a mismatch between retail loans and deposits. Most banks will not be able to exactly reconcile their retail loans and deposits. Some banks will be able to make more retail loans than they can retain in deposits and thus suffer a loss of reserves, while others will find fewer loan customers than depositors, so they will have a surplus of reserves. Banks then use wholesale markets to either 'purchase' reserves by issuing wholesale liabilities (for example, negotiable, large denomination certificates of deposit (CDs) or by borrowing central bank funds), while surplus banks will sell their excess reserves.

As discussed above, the central bank sets the overnight interbank rate. This rate then determines other short-term wholesale rates (mainly by marking up, but also by marking down) through arbitrage.

Summary

The neoclassical position is that banks leverage (create credit) when provided with new deposits, but are constrained by fractional reserve requirements. Since the central bank is supposedly able to control the monetary base, it is claimed that the central bank can control the supply of money.

Reflecting what happens in the real world, MMT demonstrates that the central bank cannot control the monetary base, because monetary policy is conducted by the central bank setting a target interbank rate and providing the right level of reserves to the banking system so that banks lend to and borrow from each other at this target rate (for more details, see Chapters 20 and 23). Second, a bank is not constrained by its reserve position in deciding whether to make a loan to a particular customer. If the customer appears creditworthy and the loan is profitable to the bank, it will make the loan and then acquire sufficient reserves by borrowing from other banks or the central bank. Thus, in contrast to the neoclassical position of deposits driving loans, MMT shows that loans drive deposits. Third, taken together, the growth in the broad money supply is driven by the demand for loans and the monetary base adjusts to the pressures that the endogenous monetary growth places on the central bank in its quest to sustain a particular policy interest rate. Hence the supply of money is determined endogenously while the price of money (short-term interest rate) is determined exogenously by central bank policy.

An example of a bank's credit creation: a balance sheet analysis

The balance sheet of a typical bank looks like that in Figure 10.2.

The entries on the balance sheet are the cheque and savings accounts. Note that they are the IOUs of banks, and hence appear as liabilities. The bank promises to convert deposits in a cheque account (and deposits in most savings accounts) into cash on demand. Banks hold financial assets in the form of loans to customers and securities (that is, treasury debt and other financial assets).

Firms in general and banks should have positive net worth, which is the difference between total assets and total liabilities. Total Assets in the left-hand column will balance with the items in the Liability column, because the latter includes net worth.

The following simplified series of balance sheets will clarify the process of credit creation by Bank A. Let us assume that Bank A starts with the very simple balance sheet in Figure 10.3, which is expressed in terms of stocks.

Figure 10.2 A typical bank balance sheet

Assets	Liabilities
Advances (loans) Securities Reserves Other assets	Cheque accounts Savings accounts Other liabilities **Net worth**

Figure 10.3 Bank A initial balance sheet

Assets		Liabilities	
Building	$200		
		Net worth	$200

Its owners have raised capital and bought the building. The owner's equity or net worth is equal to the value of the building that they have purchased. Bank A has not engaged in any banking activity yet.

Now a customer comes into the bank and says that they would like to borrow $200 to finance the purchase of a car. The bank checks their creditworthiness by asking for income tax returns, proof of assets, credit history, and so on. If the customer is approved, then the bank's balance sheet takes the form shown in Figure 10.4.

The bank just created $200 of money entries (deposits in the cheque account of the customer in return for the customer's IOU, or promise to pay $200). The bank's total assets, liabilities plus net worth, are now $400.

Before we move on to the customer's spending of their deposit, let us examine this balance sheet carefully.

Where did the bank get the money entry it created?

- A cheque account was created *ex nihilo*, that is, from nothing, by entering a number (200) in a computer ledger on behalf of the borrower. In the past, banks could also issue their own banknotes, but generally only central banks can do that now.
- The bank did not need any prior deposits, or any cash in its vault. In fact, the bank did not have any cash in its vault, nor did it have any deposits in its account at the central bank in this simplified example.
- The bank is not lending anything it has, it just creates money entries (that is, bank deposits), at will.
- Those money deposits or entries are its liabilities/IOUs.
- By creating those bank IOUs, the bank promises to:
  - Convert deposits into cash on demand;
  - Accept any of those IOUs in payment of debts owed to the bank.

The cheque account is just a legal promise to convert to cash on demand, and to accept payment in the form of the bank's own IOUs. The bank does not have to have any cash now.

The success of the banking operation (lending by accepting an IOU, and the creation of a demand deposit) depends on:

- The capacity of the customer to repay, that is, their creditworthiness. If they have problems in making timely payments on their debts, this affects the value of the bank's assets and its own income inflows and ultimately affects the net worth of the bank, the bank's capital ratio, and the shareholders' return on equity.
- The bank's capacity to acquire reserves at low cost if:
  - The customer wants to withdraw cash;
  - The bank needs to pay debts to other banks through an interbank settlement following the customer's spending (see below);
  - The bank needs to settle tax payments made by the customer to the government.

Figure 10.4 Bank A balance sheet showing loan

Assets		Liabilities	
Loan to customer	+ $200	Cheque account of customer	+ $200
Building	$200		
		Net worth	$200

If these conditions are not satisfied the bank gets into trouble; it can become insolvent or illiquid. Insolvency means that the bank's net worth falls to or below zero; illiquidity means that it cannot meet cash withdrawals or clearing. Thus, even though banks can create unlimited amounts of money

Figure 10.5 Bank A balance sheet showing purchase of car

Change in Assets	Change in Liabilities
	Cheque account of the customer −$200 Reserves due to Bank B +$200

Figure 10.6 Bank B balance sheet showing purchase of car

Change in Assets	Change in Liabilities
Claim on Bank A reserves +$200	Cheque account of car dealer +$200

deposits, they have no incentive to do so because they may become unprofitable.

What happens if the customer now pays $200 to a car dealer who has a bank account at Bank B? The balance sheets of Banks A and B look like Figures 10.5 and 10.6, respectively. (Note that we are now just dealing with the change in assets and liabilities rather than their levels.)

Bank A's liabilities in the form of the customer's cheque account have dropped by $200 through the purchase of a car, but the transaction is not confined to the reduced balances in the customer's account at Bank A and the increased balances of the car dealer at Bank B. Bank A now owes Bank B $200 and needs reserves to settle this debt, but does not have reserves. Where does it get the reserves?

The banks are required to keep reserve accounts at the central bank. These reserves are liabilities of the central bank and assets of the banks, and function to ensure that the payments (or settlements) system functions smoothly. That system relates to the millions of transactions that occur daily between banks as cheques are tendered by citizens and firms and other bodies. Without a coherent system of reserves, Bank A could easily find itself unable to fund Bank B's demands based on the cheque drawn on the customer's account and presented at Bank B by the car dealer.

Bank A will get the reserves from the source that is the least costly. It may sell assets, but in our example, Bank A only has a building so it would be very costly to get reserves that way. Bank A could sell bonds if it had any, or it could borrow reserves from other banks (domestic or foreign) or the central bank. A common way to get the reserves is to borrow from the central bank, which is the monopoly supplier of reserves. Figure 10.7 documents the latest change to Bank A's balance sheet, associated with obtaining these reserves, while Figure 10.8 shows the changes to the central bank's balance sheet.

Now that Bank A has the reserves it can settle its debt with Bank B. The changes to the two banks' balance sheets are shown in Figures 10.9 and 10.10.

The debt between the two banks has been settled. The final balance sheet of Bank A looks like Figure 10.11.

Bank A makes money as long as the interest it receives on the loan to the customer is higher than the interest it pays to the central bank on the reserves.

Figure 10.7 Bank A balance sheet showing loan from central bank

Change in Assets	Change in Liabilities
Reserve +$200	Debt to central bank +$200

Figure 10.8 Central bank balance sheet showing loan

Change in Assets	Change in Liabilities
Reserve loan to Bank A +$200	Reserve +$200

The balance sheet of Bank B is shown in Figure 10.12. We assume that Bank B had reserves prior to the cheque account of the car dealer being increased by the sale of the car to the customer.

The final balance sheet of the central bank after all transactions is shown in Figure 10.13.

Note that none of these operations involved any transfer of physical cash. It was all bookkeeping entries conducted digitally through computer networks.

Also, note that we have only shown the assets and liabilities directly related

Figure 10.9 Bank A balance sheet showing settlement of debt

Change in Assets		Change in Liabilities	
Reserves	−$200	Reserves due to Bank B	−$200

Figure 10.10 Bank B balance sheet showing settlement of debt

Change in Assets		Change in Liabilities
Claim on Bank A	−$200	
Reserves	+$200	

to our example. Of course, private banks and the central bank have many other assets and liabilities, as well as net worth on their balance sheets.

In practice, the central bank will usually not advance reserves to the bank directly in the form of an unsecured advance; instead it will ask for collateral (usually a treasury security) in exchange and will provide funds for less than the value of the collateral. So, if Bank A has a $300 bond, it surrenders it to the central bank in exchange for reserves. The central bank might only give the bank $285 if the discount rate is five per cent. The discount rate is one way in which the central bank can try to limit credit creation in the economy.

Figure 10.11 Bank A final balance sheet

Assets		Liabilities	
Funds advanced to customer	$200	Debt to central bank	$200
Building	$200		
		Net worth	$200

Figure 10.12 Bank B final balance sheet

Assets		Liabilities	
Reserves	$200	Cheque account of car dealer	$200

Figure 10.13 Central bank final balance sheet

Assets		Liabilities	
Reserve loan to Bank A	$200	Reserves	$200

Conclusion

It is insufficient and misleading to think of modern banks as 'intermediaries' that take in deposits and then lend most of them out, while retaining some fraction as reserves. Instead, we should think of banks as making loans (accepting the IOUs of borrowers) and then creating demand deposits that the borrowers can use to finance their spending. Banks mostly use reserves for clearing, that is, for settling payments with other banks, the central bank, and the treasury, and at the ATM (when cash is withdrawn). Banks obtain reserves as needed either by borrowing them from other banks or through creation of reserves by the central bank. We will explain in more detail how, and why, central banks accommodate the demand for reserves in Chapter 20.

References

Brunner, K. (1968) "The Role of Money and Monetary Policy", *Federal Reserve Bank of St Louis Review*, 50, 8–24.

Moore, B. (1988) *Horizontalists and Verticalists: The Macroeconomics of Credit Money*, Cambridge: Cambridge University Press.

Endnote

1. Orthodox economists propose that the nominal interest rate must equal a real interest rate plus expected inflation. The real interest rate is supposed to be the market-determined real return that will equate the saving supply of funds with the investment demand for funds. It is thus a real equilibrium interest rate. However, as contracts are written in nominal terms, that is, in terms of nominal interest rates, the nominal rate must include compensation for the expected inflation rate. This addition to the real rate as inflation expectations rise is called the Fisher effect (see Box 10.2). Many market participants believe this applies to bond markets. There is a strong belief that nominal yields are adjusted by markets to preserve purchasing power.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

PART C

NATIONAL INCOME, OUTPUT AND EMPLOYMENT DETERMINATION

CHAPTER

11 THE CLASSICAL SYSTEM

Chapter Outline

Learning Objectives

- Understand and assess the assumptions that underpin the classical system.
- Recognise that the continued adherence to free market principles by some commentators has its origins in the classical model.
- Realise that the return of neoclassical thinking embodied in the classical approach underlies all modern mainstream theoretical approaches.
- Appreciate the contributions of pre-Keynesian economists, notably Marx, in challenging the ideas that became known as the Classical system.

11.1 Introduction

In this chapter, we develop the 'Classical' system of employment and output determination. This approach was the conventional wisdom in British macroeconomics at the end of the 1920s and was championed by the work of Arthur Pigou in his 1933 book *Theory of Unemployment*.

The approach was attacked by John Maynard Keynes in his 1936 *The General Theory of Employment, Interest and Money*. We refer to it as the Classical system, because that is how Keynes described this body of ideas. However, it is strictly a misnomer, because the Classical economists (including Adam Smith, David Ricardo, and Karl Marx) did not use the marginal analysis that defines the British Treasury View, which is more accurately referred to as neoclassical.[1]

In Part G (Chapters 27–29), we outline the competing schools of thought about macroeconomic theory and policy, which place Classical Theory within an historical context. We review this debate from the 1930s in this chapter and the next because many of the points of contention remain relevant today. In the history of economic thought, Keynes' intervention clearly demonstrated that the neoclassical approach was deeply flawed. However, his insights were ignored in the 1970s when Monetarism emerged as the dominant school of thought (see Chapter 18). We believe it is important to understand the basis of Keynes' rejection of neoclassicism to more fully appreciate the flaws in current policy approaches, and the reasons that Modern Monetary Theory (MMT) contests much of the mainstream theory.

There are four main theoretical components to the Classical system:

- A theory of production based on the **Law of Diminishing Returns**, which links the labour market with the product market.
- The **Classical labour market**, which determines the real wage and total employment level. Under flexible real wages, the operation of market forces achieves full employment whereby everyone who wants a job can find one and every employer who wants to hire can find an available worker.
- A **theory of saving and investment** which introduces the equilibrating role of the interest rate and ensures that there can be no aggregate spending shortages and therefore no overproduction that might generate a crisis.
- The **Quantity Theory of Money** which is used to explain the general price level. The size of the exogenous money stock has no impact on the real economy but does impact nominal values.

We first explore the operation of the classical labour market and the significance of diminishing returns. This is followed by an explanation of why there may be unemployment and the determination of equilibrium output. Loanable funds theory, which is crucial to the claim that equilibrium output is associated with full employment, is explored, and the determination of the price level is then outlined.

11.2 The Classical Theory of Employment

We first consider the production function of the individual firm. This describes how much output the firm will produce for a given labour input given the stock of capital and other resources that it has at its disposal.

In the short run, the stock (and quality) of capital and other resources (such as land) are assumed to be fixed, and the only variable productive input is labour.

The Classical system is underpinned by the assertion that production is governed by the **Law of Diminishing Returns (LDR)** (see also Chapter 14). The LDR states that when a firm adds more labour input to the fixed stock of capital, output initially rises but continues to do so at a declining rate. In other words, the incremental output becomes smaller and smaller as additional units of labour are employed. The marginal physical product of labour, which is the extra output forthcoming from an additional unit of variable labour, is positive but diminishing.

Figure 11.1 shows the **aggregate production function** where real GDP (Y) is on the vertical axis and total employment is on the horizontal axis.[2] The convex shape of the production function is based on the assertion that there are diminishing returns to labour.

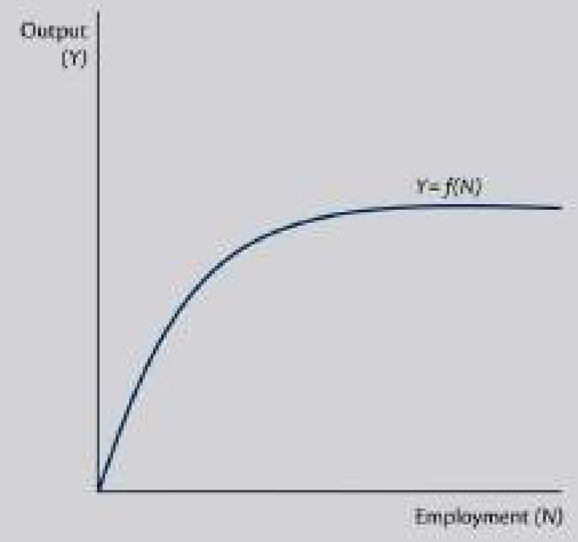
Each firm is assumed to be a profit maximiser and operates in a perfectly competitive product market, which can be summarised by the statement that each firm is too small to influence the price determined in the market for output. Thus, each firm is assumed to be a price taker and it can sell as much output as it likes at that price.

If we assume that the price level is P and the marginal physical product of labour is MP, then the value of the marginal product (that is, how much extra revenue the firm can expect to earn at each employment level) can be written as

$$VMP = P \times MP \tag{11.1}$$

where P is the price level, MP is the marginal product, and VMP is the value of the marginal product (in monetary units).

Figure 11.1 The Classical production function



The *VMP* curve is downward sloping with respect to employment because of the assertion of diminishing marginal productivity. The principle of profit maximisation determines how much a firm will be prepared to produce and how many units of labour it will employ.

Accordingly, the firm seeks to equate the return it receives from selling the last unit of output produced with the cost of producing that unit of output. This means that the firm will employ labour up to the point where the additional cost of employing an extra unit of labour exactly equals the value of marginal product (*VMP*).

The firm pays a money wage rate (*W*) to the workers. The profit maximisation rule thus means that:

$$(11.2) \quad W = VMP = P \times MP$$

which can be re-expressed as:

$$(11.2a) \quad W/P = MP$$

W/P is the nominal wage adjusted by the price level and represents the real wage paid to the workers. Thus, **a firm will employ labour up to the point where the real wage equals its marginal product of labour.**

These concepts form the basis of the Classical labour market and the Classical theory of the real wage rate and the level of employment. Accordingly, total employment is determined by the interaction between labour demand and labour supply. Once total employment is determined then the production function tells us how much output will be supplied.

The following equations define the Classical employment and output determination model:

(11.3)	Labour demand:	$N_d = f(W/P)$	$f' < 0$
(11.4)	Labour supply:	$N_s = g(W/P)$	$g' > 0$
(11.5)	Labour market equilibrium:	$N_d = N_s$	
(11.6)	Production function:	$Y = Q(N, K^*)$	

where N represents a number of hours of labour input, Y is real output, W is the money wage, P is the price level, K^* is constant capital (and other fixed productive inputs). Our specification of the production function also assumes that the state of technology remains unchanged.

The terms f' and g' are the so-called first derivatives of the respective functions (see Chapter 7 *Methods, Tools and Techniques*). Here, they tell us about the slope of the respective functions, that is, whether the relationship is an increasing or decreasing function of the real wage. The labour demand function (11.3) is downward sloping (first derivative is negative) and the labour supply function (11.4) is upward sloping (first derivative is positive) (see Figure 11.2).

Why is the labour demand function downward sloping?

Equation (11.2a) shows that a profit-maximising firm will employ labour up to the point that the marginal product equals the real wage. Given the assertion of the law of diminishing returns, the marginal product declines as employment rises. Therefore, the firm will only be prepared to employ more workers if it can pay a lower real wage rate, given that each additional worker is assumed to be less productive than the last.

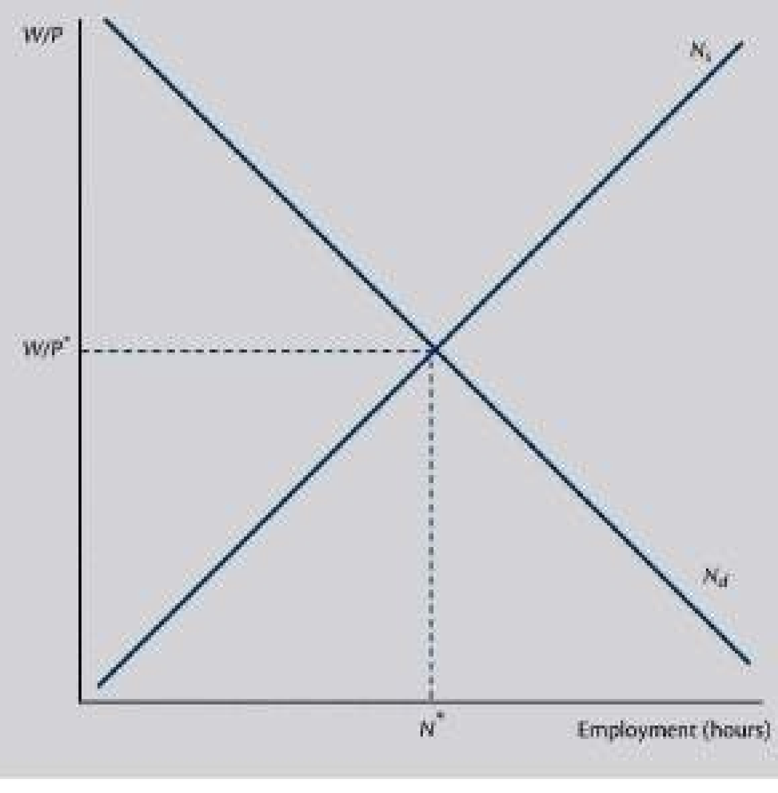
The firm will stop employing additional workers at the point where the real wage is equal to the marginal product.

The current state of technology helps explain the position of the labour demand curve. If for example, the firm invested in more efficient capital, then the production function would shift upwards and then each worker would become more productive. As a result, the labour demand function would shift outwards.

Why is the labour supply function upward sloping?

The Classical labour supply function (N_s) is based on the idea that the worker has a choice between work (which economists term a 'bad' because it undermines satisfaction or creates 'disutility') and leisure (which economists consider to be a 'good' because it adds to satisfaction or 'utility'). In this schema, work is only tolerated because it is a source of income, which allows the worker to purchase other 'goods'. The relative price mediating the choice

Figure 11.2 The Classical labour market equilibrium



between work and leisure is the real wage, which measures the price of leisure relative to income. That is, an extra hour of leisure 'costs' the real wage that the worker could have earned by working for that hour. Thus, as the price of leisure rises, the willingness to enjoy it declines.

Just as the firm is assumed to be a profit maximiser, so the worker is assumed to be a utility maximiser such that the satisfaction (utility) they derive from an extra hour of leisure is exactly equal to the satisfaction gained from the goods and services that the extra income earned from the extra hour of work would purchase.

The worker is assumed to make complicated calculations continuously by calibrating how much dissatisfaction they get from working and how much satisfaction (utility) they get from not working (enjoying leisure). The real wage is the vehicle to render these two competing uses of time compatible by allocating hours to work such that the worker maximises satisfaction.

The Classical analysis attempts to explain what happens to the number of hours of work offered by workers when the real wage changes. It decomposes the total change into two separate conceptual components: a substitution effect and an income effect.

The substitution effect refers to the impact on the worker's decision to supply hours of labour when the real wage changes. If the real wage rises, work becomes relatively cheaper (compared to leisure) and the mainstream theory asserts, via the so called law of demand, that people demand less of a good when its relative price rises. Thus, a higher real wage leads to less leisure (because time spent not working is relatively more costly in terms of the wage given up) and more work.

However, when the real wage rises, the worker now has more income for a given number of hours of work. The Classical theory then invokes the notion of normal goods (as opposed to inferior goods), for which demand increases when income rises. This is the income effect.

Classical theory assumes that like many other consumption goods which the worker might buy, leisure is a normal good. Consequently, as the real wage rises, the presence of an income effect means that the worker will demand more of all normal goods (because they have higher incomes for a given number of working hours) including leisure. Thus, the worker will work less!

Therefore within this theoretical framework, the substitution effect causes the worker to supply more(less) hours of work when the real wage rises(falls), whereas the income effect predicts that the worker will supply less(more) hours or work when the real wage rises(falls).

With these effects working in opposite directions, what determines the overall outcome? The conclusion is that despite the mathematical form of the Classical theory of employment and the pretension of analytical rigour, the theory cannot tell us unambiguously which of these two effects dominates.

The theory asserts that the substitution effect dominates the income effect over the relevant range of real wage rates, which means the labour supply function slopes upwards with respect to the real wage. The basis for this claim is the reasoning that if the labour supply function sloped downward, it might not intersect with the labour demand function.

If the negative income effect dominates the positive substitution effect above a particular real wage, then the labour supply schedule would be backward bending above this real wage, so that fewer hours of labour would be supplied. This would mean that this labour market model would not yield a coherent theory of employment determination.

We now examine equilibrium in the labour market.

Equilibrium in the labour market

Equations (11.3) to (11.5) represent a formal model of the labour market which yields both the real wage ($(W/P)^*$) and total employment (N^*) through the equilibration of labour demand and labour supply (see Figure 11.2). The Classical economists considered this level of employment to represent full employment because at that level of the real wage, all firms that wanted to employ labour could find sufficient workers and all workers who desired to work could find a firm willing to employ them.

They considered this would be the state to which the labour market would gravitate if real wages were flexible. As we will see, prior to the Great Depression, departures from this 'fully employed' state were only considered to be possible if the real wage was prevented from moving to its equilibrium level $(W/P)^*$, perhaps by trade union activity or government regulation.

The downward-sloping labour demand function and the upward-sloping labour supply function ensure that any equilibrium intersection is unique and stable so that a flexible real wage would ensure movement back to equilibrium, following a shift of either or both functions.

Thus, according to Classical theory, **the real wage is determined exclusively by labour demand and labour supply**.

As we will see in Chapter 12, Keynes showed that this key conjecture of the Classical approach was clearly false. While the nominal wage might be determined in the labour market, the real wage cannot be known until producers set prices in the product market (that is, in the shops).

11.3 Unemployment in the Classical Labour Market

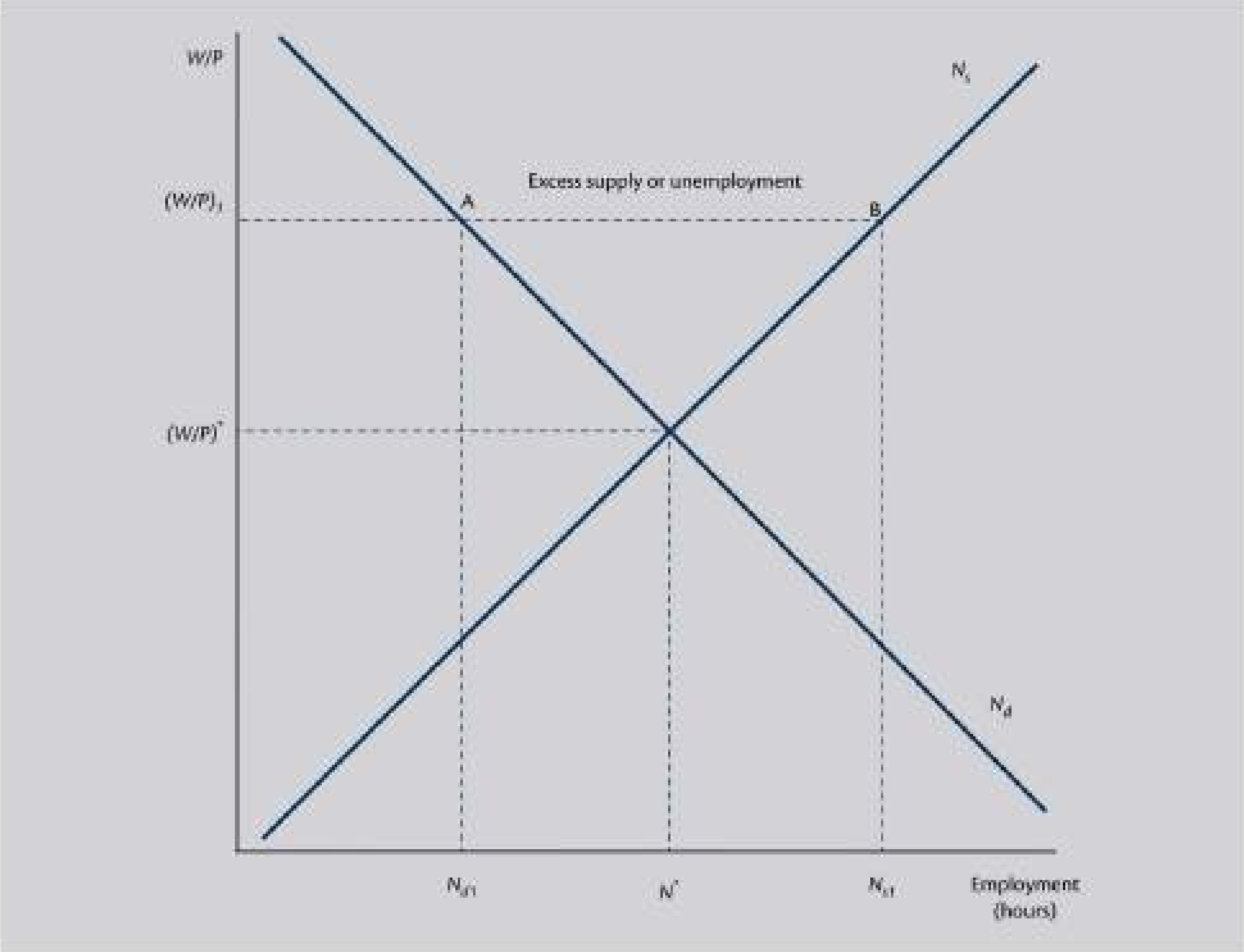
In this section we shall explain in more detail how Classical economists explained departures from full employment by the real wage being unable to achieve its equilibrium value due to institutional rigidities.

Imagine that for some reason the real wage is at $(W/P)_1$. We show this in Figure 11.3. This is a departure from the current equilibrium at the combination $((W/P)^*, N^*)$.

Workers are attracted by the higher real wage and are prepared to offer more hours of work (N_{s1}) because the price of leisure has risen and so they substitute away from the relatively more expensive good.

Firms however reduce their quantity of labour demanded (N_{d1}) in order to achieve a higher marginal product of labour, to match the higher real wage.

Figure 11.3 Unemployment in the Classical labour market



TRY IT YOURSELF

You might argue that since minimum wage legislation was not the choice of the workers but of the government, the unemployment that the Classical system generates from an excessive minimum wage would not be voluntary. What do you think? Can unemployment under the classical system be considered voluntary where it is the consequence of government legislation imposing a minimum wage above its equilibrium level?

The general rule is that the 'short side' of the market is dominant in disequilibrium situations, in which case total employment would be at A while the quantity of labour supplied would be at B. The difference (B – A) measures the **excess supply of labour** (unemployment) at $(W/P)_1$.

So unemployment can only arise in the Classical system if the real wage is above the equilibrium real wage. The Classical theory of employment denies the existence of involuntary unemployment in the macroeconomy so long as wages are flexible.

The Classical theorists believed that if the labour market were left to adjust of its own accord, the excess supply of labour would drive the price down until the labour market equilibrated at $((W/P)^*, N^*)$. Thus, the unemployment (B – A) would be considered temporary.

In other words, unemployment could not persist under Classical employment theory if real wages were flexible in both directions and thus were free to adjust to balance the forces of supply and demand.

As we have noted, the Classical economists of the 1930s recognised that institutional forces could prolong that adjustment. For example, trade unions might prevent the real wage from falling back to the equilibrium value at $(W/P)^*$. In this case, the unemployment would be considered voluntary because workers could choose to refuse union membership and offer themselves at wages below the wage floor that the union might try to enforce.

Another often-used example is the imposition of a minimum wage by the government, which might force the real wage to be higher than the equilibrium real wage. Full employment could be achieved if the government eliminated the minimum wage and allowed the equilibrium real wage $(W/P)^*$ to reassert itself.

A real wage below the equilibrium $(W/P)^*$ would result in an excess demand for labour, which would force the real wage to rise back to $(W/P)^*$, thus eliminating any imbalance.

11.4 What is the Equilibrium Output Level in the Classical Model?

Once the equilibrium level of employment is determined in the labour market, the equilibrium level of output (real GDP) is also determined given the available technology; that is, how productive the workforce is. The state of technology determines the shape and position of the production function.

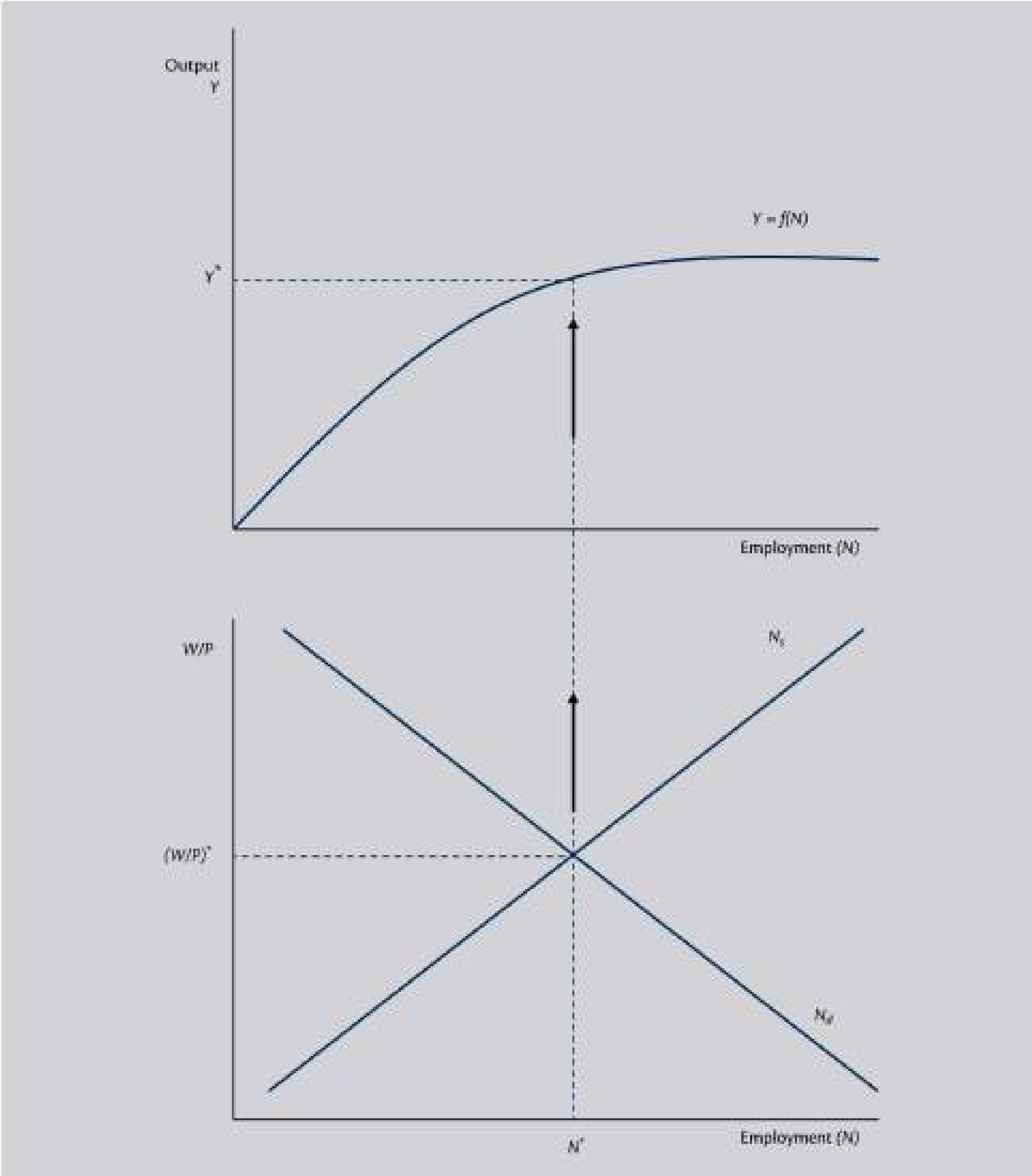
Figure 11.4 brings together the Classical labour market and the production function. The arrows in the diagram depict the Classical causality, from labour market clearing employment levels to product market output levels.

The equilibrium level of employment thus generates an output level (Y^*) which then constitutes the **aggregate supply** for the economy.

How do we know that this aggregate supply level is an equilibrium level? How did the Classical theory ensure that the labour market equilibrium was consistent with the product market equilibrium? In other words, what mechanisms ensure that the output supplied at Y^* would be purchased?

The Classical approach introduced two further theories to ensure that the aggregate demand for goods and services would always be equal to aggregate supply (Y^*). First, the Classical economists invoked what is known as Say's Law: supply creates its own demand. In short, this argument relies on the recognition that the act of producing generates income equal to the total cost of production. Further, it is argued that since the only reason to

Figure 11.4 Classical equilibrium output determination



participate in the production process is to obtain the income required to consume the output produced, it would be irrational to work to earn income that would not be spent. The problem however is that one might choose to earn income but postpone consumption to a later date. The Classical economists refer to that as investment production which will satisfy consumption later. But how can we be sure that the decision to save (postpone consumption) is equal to the decision to invest (produce goods that will be consumed later)? Only if the two were equal would it be true that supply (production today, generating income) generates equivalent demand (spending all of the income generated on either goods to consume today or investment to produce goods to be consumed tomorrow).

The **loanable funds theory** was advanced by Classical economists to demonstrate that whatever output was supplied would be demanded. More specifically, the theory asserted that movements in interest rates would ensure saving and investment are always equal at the full employment level of GDP (Y^*).

We turn to explaining that theory next, noting that Keynes demonstrated that it was flawed at the most elemental level. The reason we want you to understand this debate is because the loanable funds theory is still asserted by mainstream economists despite its inapplicability to modern monetary systems.

11.5 The Loanable Funds Market, Classical Interest Rate Determination

The Classical theory of interest rate determination provided a mechanism denying the possibility that there could be a generalised deficiency in planned aggregate spending that would result in idle productive capacity and persistent unemployment.

The Classical view agreed that specific goods and services could be 'overproduced' relative to the preferences of the consumers and firms, but at the same time claimed that rapid market adjustments would ensure that there could never be a generalised glut. The denial that generalised overproduction could occur has become known as **Say's Law** after the French economist Jean-Baptiste Say, who popularised the view.

The idea is sometimes summarised by the epithet, 'supply creates its own demand'. The logic is that by supplying goods and services into the market, producers are signalling a desire to exchange their output for other goods supplied into the market.

Assuming for simplicity a closed economy without a government sector, which is the typical depiction of the Classical system, we know that in equilibrium, the total flow of spending in the economy is equal to total real GDP and national income (Y).

Total spending is the sum of consumption (C) and investment (I) expenditure:

$$Y = C + I \tag{11.7}$$

National income is either consumed (C) or saved (S), which allows us to write:

$$Y = C + S \tag{11.8}$$

Thus, in equilibrium (total income equals total spending):

$$C + I = C + S \tag{11.9}$$

which means that the equilibrium condition is $S = I$, that is, **saving is equal to planned investment**. Thus, all consumption goods are sold and the remaining national income is equal to investment.

The Classical system considers that withheld consumption in each period is matched by investment spending, given that saving is a signal that consumers want to consume in the future. Firms are thus assumed to invest in future productive capacity to ensure that they can meet the demand that results from postponed consumption.

Assume that the economy is in equilibrium, and then there is a rise in desired saving. Consumption spending would fall in the current period and to maintain the equilibrium output level, investment would have to rise.

Ignoring the logistics of how a real world economy might quickly shift between the production of consumption and investment goods (reconfiguring machines and production processes), the **theory of loanable funds** claimed that continuous equilibration would be achieved by interest rate adjustments, which would always bring planned saving and planned investment into equality when household and firms' preferences change.

The loanable funds market is really a primitive depiction of a financial system. The interest rate (r) is a price that ensures that planned investment is equal to planned saving in any period and is market determined. Savers (lenders) enter the market to seek a return on their savings to enhance their future consumption possibilities. Firms seeking to invest enter the market to secure loans.

The interest rate that is determined in this market provides the return to households for their saving and determines the cost of borrowing funds for investment purposes.

Figure 11.5 shows the market for loanable funds. The supply of loans is derived from current saving which is assumed to be positively related to the interest rate because a rising interest rate allows savers who forgo consumption to enjoy higher levels of consumption in the future by increasing saving now. As the interest rate rises, the return on saving rises and so the supply of funds (saving) rises (and current consumption falls).

The demand for funds comes from borrowers who wish to invest in houses, factories, and equipment among other productive projects. Firms form expectations of future returns that they will derive from different projects and rank the profitability of projects given the current cost of funds. As the cost of borrowing (the interest rate) rises, the quantity of funds demanded falls because the net return on the planned projects diminishes and some become unviable. Thus, the demand for loans (investment) is negatively related to the rate of interest.

The interest rate adjusts to ensure that the supply of funds (saving) equals the demand for loans (investment). In Figure 11.5, the equilibrium interest rate is r^* per cent, which corresponds to equilibrium saving (S^*) and investment (I^*).

If the interest rate was below the equilibrium rate, then the volume of funds demanded by potential borrowers would exceed the supply of loanable funds and competition amongst borrowers would force up the interest rate. As the interest rate rises, planned saving would increase and planned investment would decline. At the equilibrium interest rate, the imbalance between supply and demand would be eliminated and planned saving would equal planned investment. The converse follows if the interest rate is above its equilibrium level.

Figure 11.6 shows the impact on the interest rate and the equilibrium level of saving and investment when households decide to consume more of their income in the current period. In this circumstance, the supply of loans (S) shifts to the left (say to S_2).

At the previous equilibrium interest rate r^*, there is now an excess demand for loans equal to the distance AB. Competition for the scarce funds drives the interest rate up and a new equilibrium is established at r^*_2 per cent, and this corresponds with equilibrium saving (S^*_2) and investment (I^*_2).

Figure 11.5 Classical interest rate determination

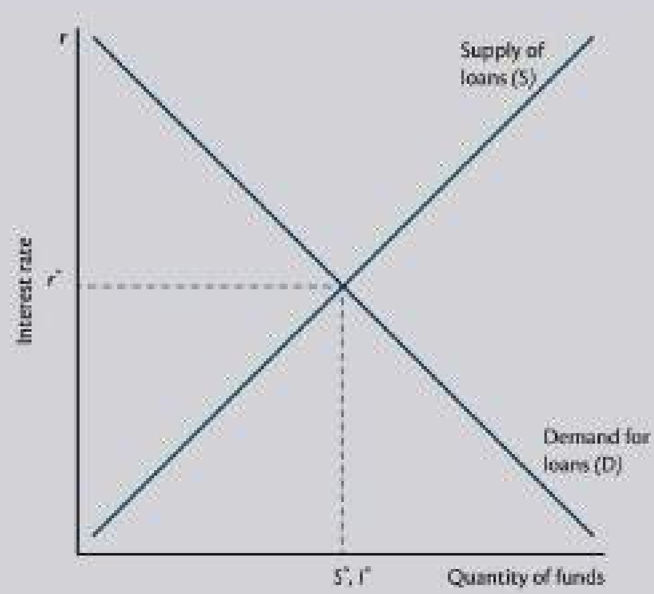
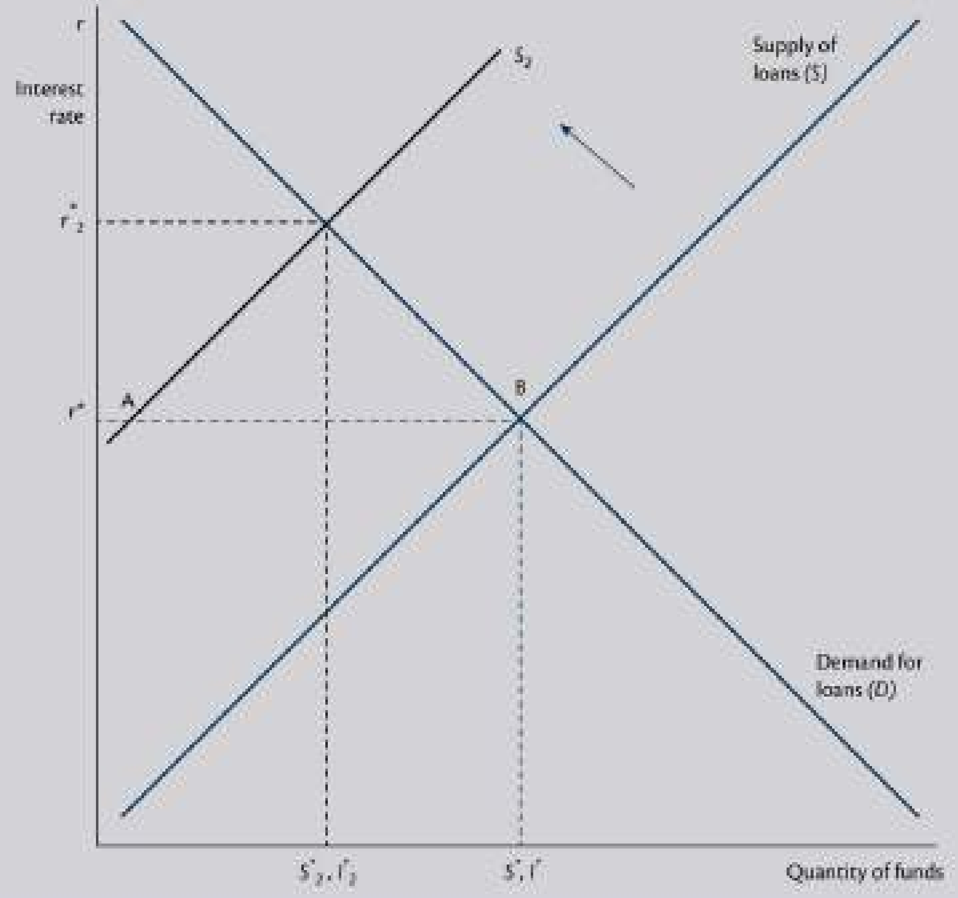


Figure 11.6 Increased desire for consumption



The real GDP level is assumed not to change because it is determined via the labour market equilibrium. The **composition** of final expenditure and output changes in the Classical system, rather than the total flow. The lower saving is offset by higher consumption, and investment contracts accordingly.

You might wonder how an economy could make these shifts quickly given that production of consumption goods and services would require a different mix of capital and labour than the production of capital goods. We will discuss that point later.

As a final observation, the Classical system considered the interest rate to be a real variable which adjusts to bring real aggregate demand into line with aggregate supply (via the loanable funds market).

Thus, the entire real side of the simple economy is explained in the Classical system without reference to money. Real GDP, national income, employment, the real wage and the interest rate are all determined in the Classical system once we know the state of technology and the preferences of households between work and leisure, and consumption and saving.

TRY IT YOURSELF

You should now be able to outline a story of what might happen within this Classical perspective should the borrowers expect stronger revenue flows from investment projects (*hint*: the demand for loanable funds would be affected) or consumers become more cautious and decide to save more of their income (*hint*: the supply of loanable funds would be affected).

As we will see in the following section, the only function that money serves in the Classical system is to determine the aggregate price level (and the inflation rate). In the economics literature, this separation in the explanation of the real side of the economy and the nominal side (price level determination) is referred to as the **Classical dichotomy**.

11.6 Classical Price Level Determination

To complete the Classical system, we consider its conception of aggregate demand, which allows the system to determine an aggregate price level. You will notice that we haven't introduced money into the model yet. The Classical system considered money to be irrelevant for the determination of employment, real output, the real wage and the interest rate.

With the output level established once the real wage and employment is determined in the labour market, and the loanable funds doctrine ensuring that real aggregate demand will be sufficient to absorb that level of output, the question that remains is how to determine the nominal variables in the Classical system.

The two key nominal variables are the aggregate price level and the money wage rate. The determination of the money wage follows once we know the price level because the real wage is just the ratio of the nominal wage to the price level.

The Classical system explains the determination of the price level by reference to the Quantity Theory of Money. Money is seen solely as a means of exchange to overcome the problem of double coincidence of wants in traditional barter models. For example, a plumber who desires the services of an electrician no longer must find an electrician who simultaneously desires the services of the plumber.

A given nominal stock of money (M_s), which is a sum of dollars, moves between individuals as various transactions are made over a period of time. The average number of times the stock of money turns over each period in the course of these transactions is called the transactions **velocity of circulation** (V_T).[3]

The product of the stock of money and the transactions velocity of circulation (M_sV_T) must therefore equal the total nominal value of transactions in a given period. Strictly the value of transactions is not equal to the value of nominal output over the same period, because as noted in the discussion of national accounts (Chapter 4), unless production is vertically integrated, there are transactions involving intermediate goods and services which are excluded in the calculation of nominal GDP. (Note that financial transactions, such as sales and purchases of financial assets, are also excluded from GDP but are included in the calculation of transactions velocity.)

Thus, to relate the money supply to GDP we need to define an income velocity of circulation, say V, which is the average number of times the stock of money turns over in the generation of national output. Thus M_sV is equal to the total nominal value of output (current GDP), which is the product of real output (Y) and the price level (P).

The Quantity Theory of Money is thus captured by the following identity:

$$(11.10) \quad M_sV = PY$$

The product M_sV is the money supply multiplied by the number of times each dollar is spent, so it will total to aggregate spending. The product M_sV is a nominal amount and represents total nominal aggregate demand in a given period, whereas the product PY represents the nominal value of aggregate supply in the same period. Obviously these are equal since they simply look at total output and income from different sides of the transaction.

The Classical system assumed that V is constant, determined by spending habits and other practices (including the time and frequency of wage and salary payments). Further, given that the Classical system assumes that the real side of the economy is determined without reference to the stock of money and that real wage flexibility ensures that full employment output will be supplied in each period (as a result of continuous labour market clearing), then Y is also assumed to be fixed in each period at full capacity output. It is clear from Equation (11.10) that if V and Y are fixed, then changes in M_s will cause changes in P.

The Classical system also assumed that the central bank controls the nominal stock of money in circulation and thus is in a position to determine the nominal value of total spending. It follows that with the real variables determined in the labour market together with the production function, the only variable that the central bank (government) can influence is the price level.

This idea resonates throughout the history of economic thought. As we will see in the next Chapter, and in Chapter 18 when we consider the Phillips curve, this idea led to the current policy resistance to the use of fiscal policy and the promotion of monetary policy as the primary stabilising policy tool.

It follows logically that if V and Y are fixed in any period, then if the central bank were to accelerate the growth in the money supply (M_s) then all that this would accomplish is an accelerating growth in the price level, which we call inflation. This theory thus provides backing for the claim that inflation is caused by lax monetary policy.

As a segue into later chapters, the Quantity Theory of Money is not a credible explanation of price level determination or inflation because in the real world V is unstable and economies clearly do not operate at full capacity output on a continuous basis. In Chapter 10 *Money and Banking*, we examined the theory underpinning the money multiplier, which the Classical system relies upon for the assumption that the central bank can control the money supply. We showed that in the real world the assumptions that underpin the operation of the money multiplier do not hold, and that the central bank is unable to control the stock of money in the economy at any point in time.

Further, as we saw in Chapter 10, the money supply should be thought of as endogenous: its quantity cannot be controlled by the central bank; instead, the quantity of money tends to expand procyclically as the demand for loans is met by profit-seeking private banks.

11.7 Summary of the Classical System

The simplified Classical system has the following features.

- Labour demand is determined by the state of technology, which is embodied in the production function, and labour supply is determined by the preference of workers for income and leisure. The real wage is the price of leisure.
- The labour market continuously clears as a result of real wage flexibility.
- Market forces which cause equilibrium in the labour market simultaneously determine the real variables in the Classical economy: real wage, employment, real GDP and the interest rate.
- The nominal side of the economy – the price level and money wage level – are determined by the stock of money that the central bank is assumed to control. The larger is the money stock relative to the given real output, the higher is the price level.
- There can be no involuntary unemployment in this system because real wage flexibility ensures that a continuous state of full employment is maintained. Any persistent unemployment must be due to rigidities that prevent the real wage from adjusting to the prevailing labour demand and labour supply conditions.
- Government policies should ensure that the real wage is flexible so that the labour market clears.
- The central bank causes inflation if it expands the money supply at a pace faster than the growth of real output, so it must exercise careful judgment in its conduct of monetary policy.

11.8 Pre-Keynesian Criticisms of the Classical Denial of Involuntary Unemployment

The 1936 publication of *The General Theory of Employment, Interest, and Money* by John Maynard Keynes is credited by many economists as providing the definitive analysis of the concept of effective demand. In Chapter 12, we will consider how Keynes attacked the Classical view that the real outcomes of the economy were determined by the full employment equilibrium achieved in the labour market via real wage flexibility.

However, the essential elements underpinning the critique of Say's Law and the modern understanding of involuntary unemployment in a monetary capitalist economy can be found in the work of Karl Marx, particularly in his *Theories of Surplus Value*.

Marx, in particular, provided a strong critique of Classical economist David Ricardo, who in his major work *On the Principles of Political Economy and Taxation* had championed the ideas of Say, which denied the possibility of generalised overproduction in a monetary economy and became known as Say's Law.[4]

In *Theories of Surplus Value*, Marx launched an attack on Say's Law. He was intent on showing that a money-using, capitalist economy was prone to economic crises (which we now call recessions) and that unemployment was an inherent tendency of such a system.

Marx challenged Say's Law as well as the 'Classical' denial that persistent unemployment could occur. He noted that Ricardo assumed that consumers had unlimited needs for commodities and having an oversupply of one good or service would be quickly overcome by increased demands for other commodities.

Marx started from the proposition that capitalists aim to accumulate ever-increasing wealth by extracting surplus value, which is production value in excess of what the workers receive in the form of wage payments. The generation of profit thus requires: (a) surplus value creation as the object of production pomotes efforts to reduce the payments to labour (limit their consuming power) and increase their productivity; and (b) the sale of commodities in markets.

On one hand, capitalists seek to repress the growth of wages and to increase work effort in order to maximise surplus value. But on the other hand, they can only realise that surplus value as money profits if they can sell all the production. The repression of wage income thus undermines their chances of achieving that. Marx readily identified this inherent contradiction in capitalism and argued that it was the precondition for crises.

Say's Law was best summarised by Ricardo (1821: 192–3) when he argued that the decision to produce is motivated by the desire to consume. In other words, a producer either consumes their own production or uses the proceeds from that production to facilitate the consumption of other goods and services. He thus considered that all production would find a buyer and there could be no generalised overproduction (supply greater than demand).

However, Marx observed that Say's Law was refuted by the fact that crises arising from insufficient spending occur in the real world. Marx saw that sale and purchase are separate actions with separate motivations. He showed that Say's Law can only hold in a barter economy which does not represent the essential features of monetary capitalism.

In barter, you may consume your own good whereas under capitalism "no man produces with a view to consume his own product" (Marx, 1861–3: Chapter 17, Section 8). A capitalist must sell and crises will occur when sales either cannot be made or are only made at prices below cost. A capitalist may have produced in order to sell (Say's Law), but if sales cannot be made how does this help?

It was clear to Marx that capitalists aim to sell to transform commodities back into money and realise profits. Consumption is not the capitalist's aim. Only the workers sell commodities (labour power) to consume.

Marx focused on the special role that money plays and demonstrated that it is more than a 'means of exchange'. It is the medium by which the exchange of commodities falls into two separate acts that are independent of each other and separate in space and time. This is the key to understanding crises in capitalism and was an insight that Keynes and others later amplified.

The break between production and the realisation of revenue from sales can occur because of this separation. For example, the existence of the chain of production and lines of credit means that a merchant may buy cloth on credit (a typical arrangement) and the farmer sells to the spinner, spinner to weaver, and so on. If the merchant cannot sell, no one in the chain is paid.

Marx's argument established for the first time that crises manifest as monetary phenomena, and that mass unemployment was not a voluntary outcome.

Marx proposed what he called the *M-C-M'* cycle. He said that modern capitalism requires that firms begin with money (working capital) to purchase productive inputs. They convert this money (*M*) into commodities (*C*), which are the goods and services that they hope to sell and realise a profit (realised surplus value) in the form of more money than they started with (*M'*). In other words, they hope to end the cycle with *M'* which equals *M* + realised surplus value. Capitalist development is thus seen as a continuous cycle in which *M* is converted into *C* with the aim of achieving *M'*.

For Marx, the M-C-M' cycle was the epitome of capitalist ambition to accumulate money as wealth. But, it is possible that people might become pessimistic and hold their money incomes in the form of saving. Money can thus be stored and spent in some future period. So, if the expectations that led the capitalists as a class to commit resources to produce C and pay incomes are not matched by the sales reality, then it is possible that they will end up with less money than they began the cycle with. In those situations, there will be a general glut of goods and services. Marx considered that possibility of crisis to be a feature of capitalist production.

Ricardo (1821) had acknowledged that while some specific goods or services could be in oversupply (for example, as consumer preferences changed), there could never be a generalised oversupply because human wants are unlimited.

Marx dealt with the question of generalised oversupply by making a crucial distinction that remains relevant in the modern debate. Overproduction has nothing much to do with absolute needs. The debate is not about whether production can outstrip needs. Instead, Marx indicated that capitalist production is "only concerned with demand that is backed by ability to pay. It is not a question of absolute over-production – over-production as such in relation to the absolute need or the desire to possess commodities" (Marx, 1861–3: Chapter 17, Section 9).

This was the first real statement of the principle of effective demand that became central to Keynes' work and which we consider in Chapter 13.

Ricardo would say that if a person wanted some shoes then they could acquire the means to buy them by producing something themselves. But he was referring to a barter economy. So why not just produce the shoes themselves? In capitalism, when there is overproduction, goods flood the market.

Marx's ideas were thus the precursor to Keynes' analysis which sought to refute the Classical notion that both the real wage and employment are determined in the labour market. In both Marx and Keynes, we see that actual employment is determined by the **level of effective demand**, that is, in the product market. While it is true that production of commodities for sale produces income to purchase them, those who receive money income do not have to spend it. Supply need not create demand for the output that would be produced.

Effective demand is the level of output at which business profit expectations are consistent with spending plans by consumers and firms. Thus, there is a limit on profitable expansion of private output. The level of effective demand places a ration on the labour market and there is no certainty that this limit will coincide with full employment, where all workers who want to work can find jobs.

In the next chapter, we turn to the critique of the Classical system provided by Keynes.

Conclusion

Our preferred pedagogy is to outline valid theory before considering the alternatives that have been proffered by orthodox economists, which we consider to be flawed. But in presenting the flawed approach before we study the Keynesian model in detail, we have departed from our usual preference. We do so in order to reflect the historical chronology of the development of these key ideas. In effect, the dominant view in macroeconomics today reflects the approach that Keynes rejected – whereas MMT draws on many of the insights offered by Keynes, whom we consider in Chapter 12. We reject the Classical system as a valid way of understanding the way the modern monetary economy functions.

Most modern orthodox macroeconomic approaches are substantially based on the Classical Theory that we analysed in this chapter and hence to varying degrees are subject to the same flaws – many of which were exposed by Keynes.

References

Harcourt, G.C. (1972) *Some Cambridge Controversies in the Theory of Capital*, Cambridge: Cambridge University Press.
Henry, J.F. (1990) *The Making of Neoclassical Economics*, Winchester, MA, and London: Unwin Hyman.

Keynes, J.M. (1936) *The General Theory of Employment, Interest, and Money*, London: Macmillan. 1957 Reprint.

Marx, K. (1861–3) *Theories of Surplus Value*. Available at: https://www.marxists.org/archive/marx/works/1863/theories-surplus-value/ch17.htm, accessed 4 April 2017.

Marx, K. (1867) *Capital: A Critique of Political Economy*, Volume I, Moscow: Progress Publishers. Available at: https://www.marxists.org/archive/marx/works/download/pdf/Capital-Volume-I.pdf, accessed 4 April 2017.

Pigou, A.C. (1933) *The Theory of Unemployment*, London: Macmillan.

Ricardo, D. (1821) *On the Principles of Political Economy and Taxation*. Available at: http://www.econlib.org/library/Ricardo/ricP5.html, accessed 3 April 2017.

Smith, A. (1776) *The Wealth of Nations*, London: Methuen. Available at: http://www.econlib.org/library/Smith/smWN-Cover.html, accessed 4 April 2017.

Endnotes

1. Here we follow current standard practice, as well as Keynes' practice, in calling this approach 'classical'. This is actually an incorrect use of the term. Strictly speaking, the classical approach is a line of thought that runs from Adam Smith's *The Wealth of Nations* (including some precursors to Smith) through to the work of Marx and Engels (and some of their followers). A distinguishing characteristic of the classical school of thought is the adoption of the labour theory of value. What is often called 'classical' is the early neoclassical theory developed in the 1870s by Walras, Menger, and Jevons. This approach abandoned the objective labour theory of value in favour of a subjective utility theory of value. Further, the neoclassical approach was developed as a direct reaction against Marx's *Capital* (published in 1867). While the true classical theory, especially as extended by Marx, was highly critical of capitalism, the neoclassical approach was created explicitly to defend capitalism from its critics. See John Henry, *The Making of Neoclassical Economics*, 1990.
2. We ignore the issues arising from aggregating employment and output across firms. Also, some participants in the debates about capital theory in the 1960s challenged the validity of using an aggregate production function in theoretical analysis. The interested reader is referred to Harcourt (1972).
3. To make the conception of velocity more clear, we can imagine that each dollar note can be spent many times over the course of the year as it passes from hand to hand as payments are made. A given stock of money can support a larger number of payments over the year if velocity is higher, as each dollar note is used more times over the year.
4. While Ricardo adopted Say's Law, he also adopted the labour theory of value that we associate with the true Classical approach. This is probably why Keynes called this the 'Classical' school although it is more properly named the neoclassical school of thought. Note also, that even though Ricardo adopted Say's Law, neither he nor the true Classical school economists believed that the principle that 'supply creates its own demand' implies that the economy operates at full employment. Indeed, the true Classical economists presumed that the economy *normally* operates with substantial unemployment. Marx called this the 'reserve army of the unemployed' and believed it was purposely maintained to 'discipline' labour. Workers would be afraid to demand higher wages and better working conditions if they knew they could be replaced easily by hiring from the pool of unemployed labour. It is the neoclassical school that combined Say's Law with the notion that flexible wages always ensure full employment, so that not only does 'supply create its own demand', but the equality of supply and demand must occur at the full employment level of production.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

12 MR. KEYNES AND THE 'CLASSICS'

Chapter Outline

Learning Objectives

- Develop an understanding of Keynes' critique of the Classical employment theory.
- Understand Keynes' critique of Classical loanable funds theory and hence recognise how the economy can reach equilibrium with widespread involuntary unemployment.
- Critically analyse the competing explanations of interest rate determination put forward by the Classical economists and Keynes.
- Develop background knowledge in preparation for studying the theory of effective demand.

12.1 Introduction

The title of this chapter is taken from a 1937 article by British economist J.R. Hicks, published in the academic journal *Econometrica*. Since then the term 'Keynes versus the Classics' refers to the debate between John Maynard Keynes, as presented in his 1936 *General Theory of Employment, Interest and Money*, and the conventional wisdom in British macroeconomics at the time represented by the work of Arthur Pigou in his book *Theory of Unemployment*. The Classical view, as we discussed in Chapter 11, claimed that mass unemployment was the result of excessive real wages.

In this chapter, we examine the arguments that Keynes presented to illustrate the inherent deficiencies of the Classical interpretations of employment, output and prices.

12.2 The Existence of Mass Unemployment as an Equilibrium Phenomenon

At the outset, the debate between Keynes and the Classical view in the 1930s centred on whether a person could become involuntarily unemployed. The Treasury View denied the existence of **involuntary unemployment** and argued that fiscal policy (government spending and taxation) could not enhance national prosperity by creating employment.

In his 1929 budget speech, delivered 25 April (Hansard, 1929), the British Chancellor of the Exchequer, Winston Churchill outlined this view very clearly:

> *The orthodox Treasury view, and after all British finance has long been regarded as a model to many countries, is that when the Government borrows in the money market it becomes a new competitor with industry and engrosses to itself resources which would otherwise have been employed by private enterprise, and in the process it raises the rent of money to all who have need of it.*

The Treasury View is clearly aligned with the loanable funds theory, which we outlined in the previous chapter, because it assumed that there could never be a situation in which resources, especially labour resources, were left idle. And since all resources are fully employed, any attempt by government to use more resources would simply divert already-employed resources from some other use. In Chapter 21, we consider the modern debates in macroeconomics in more detail, and you will readily see that the so-called **crowding out** objections against the use of fiscal policy, that are still entertained in the current era, are based on this Treasury View. In fact, the current attack on the use of fiscal policy to increase employment growth reflects the conservative position put forward in the 1920s and 1930s, which we have characterised as Classical employment theory.

Churchill also denied that fiscal policy would deliver lasting employment gains: "very little additional employment and no permanent additional employment can in fact and as a general rule be created by State borrowing and State expenditure".

As we saw in Chapter 11, the Classical approach led to the conclusion that real wage flexibility would guarantee the achievement of full employment. Keynes disputed this reasoning and outlined a new approach to the labour market, one which provided an explanation for mass unemployment that was independent of whether money wages were flexible or not. At the time, this contention was revolutionary because it challenged the intellectual supremacy of the conservative economists who held sway in government.

Keynes demonstrated that mass unemployment arises because of the **systemic failure** to create enough jobs, which leaves individuals powerless to improve their own circumstances.

As a precursor to understanding the way in which Keynes developed his new theory, it is important to consider two broad concepts of **equilibrium** that can be found in the literature.

One way of construing the concept of equilibrium is to think of it as the position at which the economy will remain if no disturbances to supply or demand occur. Under Classical theory, equilibrium is a position of rest, but it is also commensurate with market clearing, that is, where supply equals demand.

Unemployment therefore in Classical employment theory is a **disequilibrium** phenomenon (there is an excess supply of labour) which will be eliminated since real wage flexibility will restore the market clearing equilibrium.

The concept of involuntary unemployment that Keynes introduced into the literature was consistent with the concept of equilibrium being a state of rest. But it was in sharp contradistinction to the idea that equilibrium also requires equality between demand and supply. In other words, he advanced an argument that **mass unemployment is an equilibrium state** that the capitalist monetary economy tends towards, and in which it can remain indefinitely, in the absence of government intervention.

Therefore, while the Classical approach viewed unemployment as a temporary disequilibrium state, which would be soon corrected as real wages adjusted to remove the demand and supply imbalance, Keynes saw unemployment as being an equilibrium state, which persists unless **effective demand** is increased. We consider the concept of effective demand in detail in Chapter 13. For Keynes, making wages more flexible cannot solve the problem of unemployment, and might even make the problem worse. Identifying unemployment with labour market disequilibrium leads to an incorrect diagnosis and a counterproductive policy recommendation. According to Keynes, we must start from the recognition that involuntary unemployment is consistent with equilibrium given the level of effective demand. The problem is with the level of effective demand, not with the level of wages.

This distinction also influenced how Keynes defined full employment. For the Classical economists in the 1920s, full employment occurred whenever labour demand and labour supply were equal, irrespective of the associated level of employment. They denied the existence of unemployment so that the number of people of working age not in employment reflected voluntary choices by these people not to work.

For Keynes, the achievement of full employment required that effective demand (total spending) is sufficient to ensure that there are enough jobs on offer to match the willing labour supply at the current money wage level. It is a state that the capitalist system might achieve (as a special case) but there is no general tendency within the dynamics of the system to move the economy to this state.

12.3 Keynes' Critique of Classical Employment Theory

In the *General Theory of Employment, Interest and Money* (1936) Keynes sought to show that the Classical theory failed to provide a satisfactory explanation for the existence of mass unemployment. He also rejected the principal policy proposed by the Treasury View that money wages should be cut to solve unemployment.

Keynes argued that even if real wage flexibility was possible, the economy could still remain in a state of mass unemployment. In other words, he sought to demonstrate that the existence of mass unemployment was not related to whether real wages were flexible or not. As a matter of policy, which we will explain later, Keynes thought that the existence of money wage rigidity and the institutions that supported it was preferable. But he was at pains to show that mass unemployment is not caused by that institutional structure.

In Chapter 2 of the *General Theory*, Keynes introduced what he called the 'two fundamental postulates' of Classical economics (that is, the Treasury View):

i. The wage is equal to the marginal product of labour ...
ii. The utility of the wage when a given volume of labour is employed is equal to the marginal disutility of that amount of employment. (Keynes, 1936: 5)

REMINDER BOX

You should ensure that you can relate these 'postulates' to the discussion in Chapter 11. The first postulate describes the Classical labour demand theory by which profit-maximising firms hire workers up to the point where the real wage paid to the last worker hired is exactly equal to the marginal productivity of that worker.

The second postulate relates to the Classical theory of labour supply that workers choose between work and leisure. The former provides income, which allows the worker to derive 'utility' from the purchase of goods and services, but also is considered to be bad ('a disutility') because it diverts workers from enjoying leisure.

The price of leisure is the real wage and each worker ensures that the number of hours of work they supply is such that the satisfaction (utility) derived from an extra hour of leisure (not working) is exactly equal to the satisfaction gained from an extra hour of work and the goods and services that the extra income would purchase.

While Keynes ultimately showed that the real wage and total employment level are not determined in the labour market, as held by the Classical theory, he accepted the first postulate for the sake of argument. His main motivation was to divert the focus of his critique onto the supply side of the labour market, which he considered to be the source of the Classical failure to understand how involuntary unemployment could arise as a normal tendency of a capitalist monetary economy.

Keynes had two objections to the second Classical postulate. The first concerns the way workers see real wages and money wages. Thus, he believed this postulate does not accord with the real world behaviour of workers, and suggested that workers behave in an asymmetric way to real wage reductions, depending on whether they are motivated by money wage reductions or a rise in the general price level.

He argued that workers would withdraw their labour services if money wages were cut (and real wages fell) but would not respond in this way if an equivalent real wage cut resulted from a rise in the price level. In other words, workers **do not** necessarily withdraw their labour services when the real wage falls. It depends on what causes that reduction. Keynes argued that the Classical school failed to make this distinction.

The clue to understanding his argument was to note that the Classical model assumed that labour supply is a function of the real wage exclusively (see Figure 11.2, p. 167). However in the real world, workers are also concerned with the **level of money wages**, as well as the purchasing power equivalent of the real wage.

The Classical economists' response to this critique was to claim that it is irrational or illogical for workers to suffer from what they called 'money illusion'. That is, why would workers care about the nominal value of their wage? Surely, it is only the real wage that matters because their decision to supply labour is to acquire goods and services from the income they earn?

Keynes' response was telling. First, while a rise in the general price level affects all workers, money wage cuts would be typically applied to only certain segments of the workforce (where unemployment is concentrated) and alter the position of those workers relative to all other workers. Research confirms that workers are influenced by their **relative** place in the wage structure, given that wages are one way in which we measure social status (see, for example, Bracha et al., 2015). One's money wage is more visible to others than the more ambiguous concept of the real wage. At parties and other social milieu, we informally judge each other by the income levels that we receive. These are concerns that the Classical model ignores.

If workers in a particular industry were to accept a money wage cut when that industry was enduring a downturn in demand for its output, then they would be downgrading their position in the wage structure. They would also form the view that their relative position in the wage structure would not be reversed when the economy improved again. As a result, these workers will resist a money wage cut. However, they will not necessarily resist a real wage cut (of the same magnitude) arising from a rise in the general price level because this would impact all workers and their relative positions in the wage structure would be maintained. They would all be worse off, but not in relative terms.

While the Classical approach ignored the richness of social institutions, thinking of them as ephemeral rigidities standing in the way of competitive outcomes which would disappear under the force of competition in the long run, economists like Keynes understood their significance. This made it easy for him to understand that it is perfectly rational for a worker to be concerned about social standing (and hence wage relativities).

Keynes also understood that trade unions are important institutions in a capitalist economy. They try to protect the place of workers in the income distribution by resisting any cut in their money wages, but industrial action would not be undertaken each time a rise in the cost of living occurred.

The second reason why workers would resist cuts to money wages relates to the financial arrangements that workers enter into in the normal course of their lives. A major commitment that many workers take on is the purchase of a home. Also, workers use credit to smooth their consumption expenditure over time. These contractual commitments are always specified in nominal (that is, money) terms. For example, a worker with a home mortgage (or a student loan) must pay a certain quantity of dollars per month to service that commitment. In other words, the solvency of the worker is a nominal concept. If workers cannot get sufficient money income each period to service their nominal contractual commitments, then they are in trouble.

In this context, if the general price level rises and the real value of their money wage declines, for a time they can change their budget allocations (perhaps eliminate some non-essential items of expenditure) and still maintain their nominal contractual obligations. At some point this becomes impossible, but within the usual variations in the real wage this is how households cope.

However, if the real wage was to be adjusted via reductions in the money wage in response to unemployment, workers might find they do not have enough money income in a period to service their contractual obligations and they would then have to default and face insolvency. Clearly, it is rational to resist that eventuality and thus workers care not only about the real wage they are able to earn, but also the level of money wages that they receive.

However, Keynes did not consider these institutional objections to be fundamental to the theoretical veracity of the Classical employment model. His second major objection, however, went to the heart of his theoretical attack on the Classical explanation of unemployment. He argued that a money wage reduction, even if agreed by workers, may not necessarily flow through to a fall in the real wage.

The Classical model characterises the interaction between labour demand and supply as being mediated by movements in the **real wage**, yet in the real world it is the money wage that is agreed in the labour market.

Thus, workers apply for jobs, which specify a certain money wage that will be paid. In some cases, workers negotiate the money wage they are prepared to accept. The point is that the so-called labour market contract that leads to a worker taking a job with an employer is followed by a money wage being paid periodically to the worker. It might be $15 per hour or $80,000 per year or whatever.

To argue that unemployment is voluntary and can be solved by a reduction in the real wage assumes that workers have volition and can engineer the appropriate real wage cut by accepting lower money wages. Keynes disputed the Classical claim that the bargained money wage directly determined the real wage. Accordingly, he believed that even if we were to accept the Classical logic that lower real wages would increase employment, workers as a class were powerless to deliver real wage cuts through the money wage agreements they made with firms.

Keynes' logic was that if a money wage reduction occurred, it was likely to lead to lower prices because marginal costs, which are largely determined by money wages, would be lower (ignoring shifts in productivity due to issues relating to workforce morale). Imagine that money wages fell by five per cent and the price level fell by five per cent, then the real wage would be unchanged and the level of unemployment would be practically the same. Thus, Keynes' fundamental objection to Classical employment theory was that even if workers agreed to work for lower money wages, this did not necessarily guarantee a real wage cut.

Before we consider what this means in terms of Keynes' own construction of the labour market, we should note that by accepting the marginal productivity rule (the first Classical postulate) that the real wage was equal to the marginal product of labour in the *General Theory*, he was agreeing with the proposition that for employment to rise, the real wage had to fall. As we will see, the causation that Keynes invoked to explain that association between employment and the real wage was different to the Classical theory. But in his refutation of the Classical employment theory, Keynes considered that firms are always 'on' their demand for labour curve and aggregate demand fluctuations shift employment up and down that curve.

Thus, he argued that the best way to engineer reductions in the real wage is not to try to tinker with money wages (for all the reasons we have just considered) but rather to generate some inflation by stimulating aggregate demand. He thought that as the demand for goods and services rose, firms would expand production, encounter increased marginal costs and so push up the profit-maximising price level.

However, his point was that with a rigid money wage, unemployed workers would be prepared to supply labour to firms if they offered more jobs, even though the real wage would fall as a result of a rise in the general price level. Unemployment is thus driven by the lack of spending on goods and services rather than an excessive real wage. But given a nominal wage, it is true that the real wage might fall as employment rises if what Keynes called the elasticity of production is less than one, so that firms respond to an increase of demand through a combination of increasing production and raising prices. We will return to this argument in the next section.

BOX 12.1 IS THERE AN INVERSE RELATION BETWEEN EMPLOYMENT AND REAL WAGES? A CRITIQUE OF THE FIRST POSTULATE

Keynes rejected what he termed the Second Postulate (which related to the labour supply curve) but he accepted the First Postulate (linking employment demand to the real wage). We consider here whether we should accept a negative relationship between employment and real wages.

The fact that both major theories – Classical and Keynesian – at this stage agreed that there would be an inverse relationship between employment and the real wage makes it difficult to empirically examine the veracity of either theory, even though the underlying causality that creates the relationship is very different. In other words, if we observed a fall in real wages accompanying a rise in employment, which theory would be correct?

As an historical note, by 1939 Keynes had changed his view about this inverse relationship. Two separate studies persuaded him that his earlier views on marginal productivity theory were not supported by the evidence. One was published in 1938 by the American economist John Dunlop, and the other, a year later, by Canadian economist Lorie Tarshis. These articles demonstrated that there was no definitive inverse relationship between real wages and employment, which meant that the idea that a reduction in unemployment could be accomplished by driving up the general price level to deflate a fixed money wage was unsustainable.

Keynes responded to these articles in his own 1939 *Economic Journal* article by saying that the research presented: "clearly indicate[s] that a common belief to which I acceded in my 'General Theory of Employment' ... needs to be reconsidered ..." (p. 34). Thus, it was likely that "the falling tendency of real wages in periods of rising demand" (Keynes 1939: 40) was contrary to the real world evidence. He suggested that this strengthened the argument that effective demand, rather than real wages, was the crucial determinant of employment.

In his attempt to introduce greater conceptual clarity in his debate with the Classical economists, Keynes also introduced the rather difficult concept of **involuntary unemployment** which follows directly from the notion of full employment as a maximum employment level that satisfies the preferences of the workers for work. The idea is that there would be no involuntary unemployment if all persons willing and able to work could find jobs at the prevailing wages.

In Chapter 2, Keynes (1936: 15) said that: "Men are involuntarily unemployed if, in the event of a small rise in the price of wage-goods relatively to the money-wage, both the aggregate supply of labour willing to work for the current money-wage and the aggregate demand for it at that wage would be greater than the existing volume of employment."

Apart from rendering the gender non-specific, the same definition holds today. But what does it mean? Read literally, it invokes a mental experiment in which the real wage is reduced as a result of price inflation being greater than the growth in money wages, and one observes firms seeking to hire more labour and workers supplying more labour hours, even though the real wage is lower than before.

You will appreciate from our discussion that this definition was predicated on his view that real wages would be lower at higher levels of employment, a view he subsequently abandoned. The definition is however helpful in summarising the point that the money wage level is not the problem when we are trying to understand unemployment.

It also focused Keynes' attention on the second postulate of Classical theory relating to the claim that macroeconomic labour supply was an increasing function of the real wage and the equilibrium employment level is determined by the equality of labour demand and labour supply. If workers supplied more labour even though real wages had declined, then this observation seriously compromised these two Classical claims and largely negated that theory of employment.

However, in the strategic context of his debate with the Classical economists, Keynes' definition maintained the focus on the labour market, which allowed his essential insight that a credible theory of employment should be based on the principle of effective demand (that is, the product market) to be somewhat obscured.

12.4 Involuntary Unemployment

In his first definition of involuntary unemployment, Keynes conducted a mental experiment which entailed examining the impact on the labour market of a cut in the real wage (see the box above). Here we develop Keynes' second definition of involuntary unemployment, based on his arguments outlined in Chapter 3 of the *General Theory*.

Accordingly, we might conclude that involuntary unemployment exists:

> **If a rise in effective demand in the product market leads to a rise in employment independent of any changes to wages.**

You will appreciate that this definition of involuntary unemployment highlights the fact that labour demand and hence employment is driven by shifts in aggregate demand. In other words, if firms need more employment to satisfy higher expected demand, then there is involuntary unemployment if they can increase employment without offering higher wages because there are unemployed workers willing to work at the going wage.

On the other hand, if aggregate demand rises and no further increases in employment are observed, then the economy would be at full employment. Once again this allows us to define full employment in terms of a number of jobs rather than in terms of some particular real wage. Unlike the Classical theory of employment, which considers full employment to be determined in the labour market, our definition clearly relates the concept of full employment to spending developments in the goods and services market.

To better understand this concept, let us think how we might reconsider the macroeconomic labour market. The labour supply function that is consistent with this concept of involuntary unemployment is quite different to that presented in the Classical model. Instead of being an increasing function of the real wage, based on the assertion that workers face a trade-off between labour and leisure which is mediated by movements in the real wage, Keynes argued that workers supply their labour based on the current money wage level.

Workers face considerable uncertainty about the price level; something that they cannot control. The price level is not set in the labour market and workers have imperfect information about the current price level (and have even more uncertainty about the path of the price level over the period of the contracted money wage).

The relevant labour supply function is thus written as:

$$(12.1) \qquad N_s = f(W, P^e) \qquad if\ N_s < N^*$$

which says that aggregate labour supply (N_s) depends on the money wage level (W) and the expected price level (P^e) up to full employment (N^*).

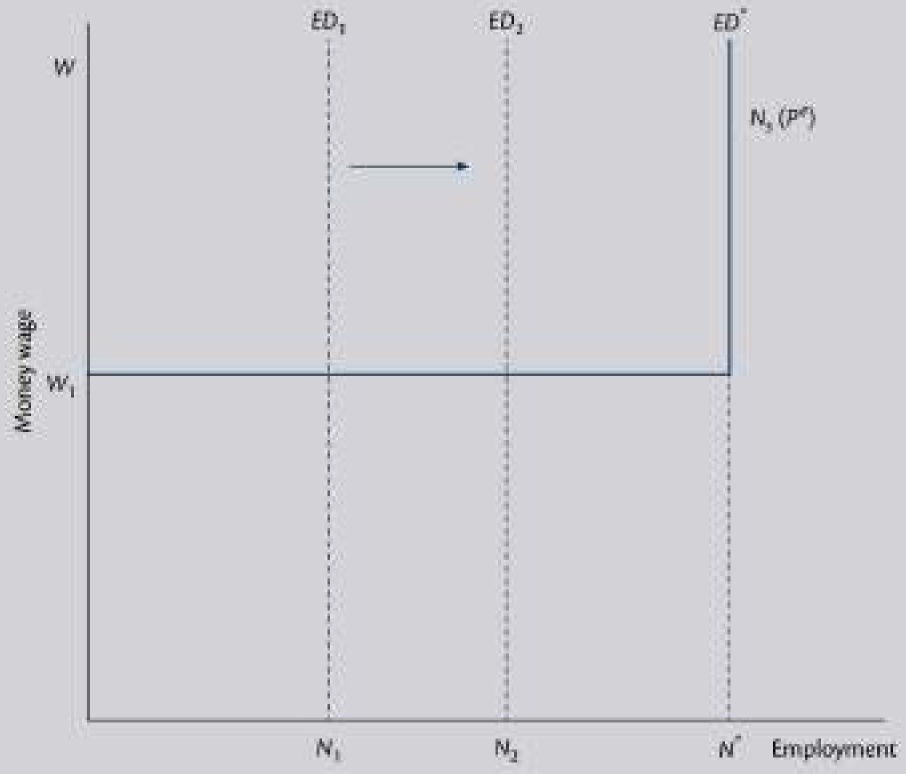
Figure 12.1 shows the Keynesian labour supply function. In contrast to the Classical labour market depicted in Figure 11.2, the money wage (W) rather than the real wage (W/P) is determined in the labour market. Workers prefer higher real wages to lower real wages, but when they enter the labour market it is the money wage they agree on with the employer, not the real wage.

Full employment occurs at N^* and beyond that level we might assume that there would be no further increases in labour supply. In the real world, it is likely that rising money wages at the current price level might attract extra workers into the labour market, people who would not usually wish to work. In other words, the vertical segment in the labour supply might slope slightly to the right to account for rising labour force participation induced by higher wages.

Up to N^*, at the current money wage level (W_1), workers will be willing to supply whatever labour is demanded.

The vertical lines, ED_1 and ED_2, reflect macroeconomic demand for labour curves that are driven by the different levels of effective demand, which is determined in the goods and services market (that is, outside of the labour market). We develop a formal version of the macroeconomic demand for labour curve in Chapter

Figure 12.1 Keynesian aggregate labour supply function



TRY IT YOURSELF

Assume that the economy is operating at an effective demand level, ED_1 and employment is at N_1. How much involuntary unemployment is there at that level of economic activity?

14. For now, the *ED* lines can be thought of as constraints imposed on the labour market, with employment adjusting at the current money wage levels to shifts in these constraints that are associated with changes in effective demand.

The answer to the question raised in the box above is the distance $N^* - N_1$, because if effective demand were at ED^*, the labour supplied would increase to N^*. To see that more clearly imagine that aggregate demand increased and the economy moved to a higher level of effective demand, ED_2. With no change in the money wage rate (or the price level), employment would increase to N_2 and unemployment would **fall** by $N_2 - N_1$.

This reasoning allows you to understand that the macroeconomic level of unemployment is not exclusively determined in the labour market, and variations in unemployment are driven by variations in effective demand.

In conclusion, Classical theory claimed that full employment was guaranteed by the real wage being flexible in response to market forces, that is, the demand and supply of labour. On the other hand, Keynes emphasised the role of effective demand in determining output and employment.

12.5 Keynes' Rejection of Say's Law: The Possibility of General Overproduction

Refresher: the loanable funds market

The Classical theory of interest underpins Say's Law and the claim by Classical economists that there cannot be a situation where aggregate demand is insufficient to absorb all production. The contention is that when firms choose to supply goods and services they produce output and income. That income must be either consumed or saved.

In Classical theory, the interest rate ensures that the income that is not consumed in each period (that is, which is saved) is also equal to aggregate investment (the creation of new productive capacity). This is the **loanable funds doctrine**, which we considered in Section 11.5.

The doctrine is based on the claim that the adjustment of the interest rate brought planned investment (the demand for loans) into line with planned saving (supply of loans). The hypothesised movements in interest rates thus ensured that there could never be an enduring excess demand or supply of loans, which then meant that aggregate demand for goods and services would always adjust to movements in aggregate supply. So, full employment was guaranteed by the combination of the adjustment of:

- The real wage to correct any imbalances between the demand for and the supply of labour; and
- The interest rate to achieve a composition of final output (between current and future goods) that is always consistent with the aggregate spending (consumption and investment).

As we will see, there is also a major issue of what market signal is generated by a decline in consumption. The implication is that rising saving signals a desire to consume more in the future and firms produce capital goods to meet that demand in the future. But what kind of consumption goods will be demanded in the future and when? These are major challenges for the Classical economists and were part of Keynes' attack on the Classical theory.

Keynes' critique of the loanable funds doctrine

Keynes was writing at the time of the Great Depression. Investment had fallen, as had national income, as economies plunged into crisis with rising unemployment. The flow of saving also fell in proportion with the decline in national income. While all these variables seemed related, they were largely disconnected from movements in interest rates and the key propositions of loanable funds theory appeared to lack foundation.

Keynes thus rejected the Classical belief that the household decision to save is determined by the preferences for current and future consumption mediated by the interest rate (the price at which consumers traded current consumption for future consumption). Instead, he considered **aggregate saving to be a positive function of national income.** This would mean that when national output and income rise, aggregate saving would rise. The amount of extra saving per dollar of additional disposable income is called the **marginal propensity to save (MPS)**. If the MPS = 0.20, then households would save 20 cents of every extra dollar of disposable income that they receive. This would be a residual after the households had consumed 80 cents of every extra dollar of disposable income received.

The interest rate might have some influence on saving, but Keynes considered national income to be the dominant factor determining the aggregate level of savings in any period. The other consideration is that investment spending is a component of aggregate demand, which in turn, drives total national income in each period.

Taken together, these insights undermine the concept of a loanable funds market as conceived by the Classical economists. There could not be independent saving and investment functions brought together by movements in the interest rate because investment drives income, which influences saving.

To conclude, Keynes found that once we realise that investment and saving functions are both functions of national income, the theory of interest rate determination provided by the loanable funds doctrine fails because it provides no way of knowing how far the investment and saving functions might shift at different levels of national income. (See Box 12.2 for a graphical exposition of this reasoning.)

BOX 12.2 GRAPHICAL EXPOSITION SHOWING SAVING AND INVESTMENT ARE NOT INDEPENDENT

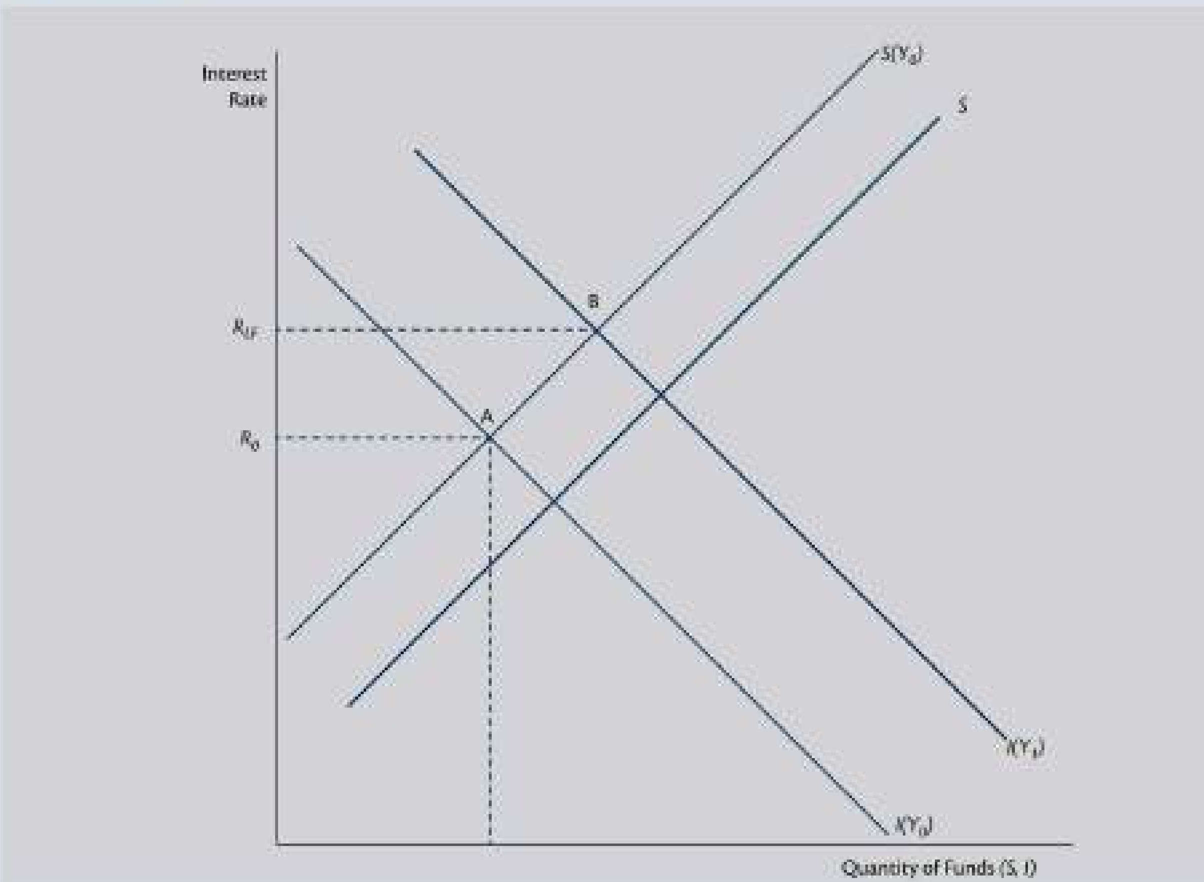
In Chapter 14 of his *General Theory of Employment, Interest and Money*, Keynes (1936: 180) produced a diagram to illustrate his contention that the interdependency of savings and investment meant the loanable funds doctrine was a 'nonsense theory'.

Keynes wrote (1936: 179–80):

> *The independent variables of the classical theory of the rate of interest are the demand curve for capital and the influence of the rate of interest on the amount saved out of a given income; and when (e.g.) the demand curve for capital shifts, the new rate of interest, according to this theory, is given by the point of intersection between the new demand curve for capital and the curve relating the rate of interest to the amounts which will be saved out of the given income. The classical theory of the rate of interest seems to suppose that, if the demand curve for capital shifts or if the curve relating the rate of interest to the amounts saved out of a given income shifts or if both these curves shift, the new rate of interest will be given by the point of intersection of the new positions of the two curves. But this is a nonsense theory. For the assumption that income is constant is inconsistent with the assumption that these two curves can shift independently of one another. If either of them shift, then, in general, income will change; with the result that the whole schematism based on the assumption of a given income breaks down ... In truth, the classical theory has not been alive to the relevance of changes in the level of income or to the possibility of the level of income being actually a function of the rate of the investment.*

Figure 12.2 shows the situation graphically. It is similar to Keynes' diagram, but the axes have been reversed in order to maintain continuity with Figure 11.6.

Figure 12.2 The interdependence of saving and investment



Relating Keynes' lengthy argument to Figure 12.2 is straightforward. The family of saving functions, $S(Y_0)$, $S(Y_1)$, and $S(Y_2)$ are drawn for different interest rates.

The family of investment functions, $I(Y_0)$, $I(Y_1)$, and $I(Y_2)$ show that when investment is higher, national income rises.

The loanable funds equilibrium, say at Y_0, would be at the intersection of $S(Y_0)$ and $I(Y_0)$, generating an interest rate of R_0 (Point A).

Loanable funds theory did not allow for changes in national income to impact on saving in this way. Accordingly, imagine investment moves from $I(Y_0)$ to $I(Y_1)$. We know that the shift in investment will drive up national income because investment is a component of aggregate demand. The loanable funds doctrine would argue that the shifting investment function would move along the saving function (S_0) and the interest rate would increase to R_{LF} (Point B).

The rise in the interest rate would be necessary, according to loanable funds theory, to eliminate the excess demand for funds at R_0 arising from the rise in investment.

Keynes argued that if saving is a function of national income then the Classical approach fails to describe the adjustment in interest rates. He said (1936: 181) that: "the above diagram does not contain enough data to tell us what its new value will be; and, therefore, not knowing which is the appropriate ... [saving function] ... we do not know at what point the new investment demand-schedule will cut it."

The increase in income also shifts the saving curve to the right although we do not know whether this will be to S_a or S_b, for example. It all depends on how sensitive the saving relationship is to national income changes.

What happens to interest rates, according to Keynes, depends on the slopes of the saving and investment lines and the amount they shift. We show in Figure 12.2, that the interest rate would rise if the saving curve shifted to S_a but fall if the saving curve shifted to S_b. The important point is that we do not know how the interest rate will change.

In other words, the loanable funds doctrine could not explain movements in interest rates and a new theory was required. It was at this point that Keynes proposed the concept of **liquidity preference** as the foundation of his theory of interest rates.

The fundamental flaw of Classical theory was that the level of national income was considered to be constant at the full employment level. Employment theory was based on an abiding faith that real wage movements would ensure that the demand for and supply of labour were in balance so that full employment would be constantly achieved.

It was obvious to Keynes that national income and employment are not constant and certainly during the Great Depression, the mass unemployment demonstrated that enduring departures from full employment were a basic feature of the economic system. With investment and saving also moving with shifts in national income, Keynes concluded that a theory of interest rates had to be found elsewhere.

Liquidity preference and Keynes' theory of interest

Liquidity preference was a concept introduced by Keynes to provide the basis for an alternative theory of interest to that found in the loanable funds doctrine. He argued that a theory of interest has to be grounded in an understanding of the way in which money and financial markets operate; in particular, the way in which people and firms adjust their wealth portfolios of money and bonds.

Recall that the Classical theory of interest considered the interest rate to be a real rather than a monetary variable. Savers are paid interest as a reward for abstaining from spending, thereby forgoing the consumption of real goods and services now and allowing these resources to be invested in order to build productive potential.

How can borrowers afford to pay interest? For the Classical economists, the answer was easy. By using the savings of households, firms could become more productive in the future through the accumulation of capital and the interest paid came from the extra real goods and services that the economy could produce.

Thus, savers enjoy a premium in terms of higher future consumption of real goods and services made possible by the increased investment that the borrowing allowed. Saving is a real act – the forgoing of consumption of real goods and services now – as is investment – the construction of capital leading to increased goods and services in the future.

For Keynes however the interest rate is a **monetary** rather than a real variable, and is an important input into a plethora of financial decisions that people make in determining the composition of their wealth portfolios.

In contrast to the loanable funds doctrine, which considers that the rate of interest determines the flow of saving and investment in the economy, Keynes believed that the level of effective demand determines saving via the propensity to consume. Once aggregate demand is known, the level of consumption is determined by the resulting national income level.

Keynes (1936: 166) said that the "propensity to consume … determines for each individual how much of his income he will consume and how much he will reserve in some form of command over future consumption". The 'command for future consumption' is what we call saving and is a residual from the act of consumption, which is largely driven by movements in national income (and taxation). The interest rate does not determine this flow.

Recall that the Classical view is that saving is postponed consumption. Essentially what these economists had in mind is that when one decides to save, one is actually ordering goods to be consumed tomorrow. This order for tomorrow's goods is placed with firms that undertake investment so that they will have the productive capacity in place to produce those goods at some future period; hence there is no reduction of effective demand resulting from decisions to save.

Keynes' argument was that while individual saving clearly signals an intention to consume in the future rather than today, and thus reduces current demand, there is no information forthcoming to producers as to when and where the future spending will occur. There is thus no reliable market signal sent to firms when households save, other than that they are not spending. There is no information in the rising inventories to tell firms when households **will** spend those savings in the future and **what** goods and services they will require. This is why 'postponing consumption' depresses demand; and rather than investment rising to fill the demand gap, investment itself might be discouraged due to the fall of effective demand. For this reason, Keynes rejected the Classical theory of the loanable funds market and argued that the interest rate does not equilibrate saving and investment.

So what impact does the interest rate have?

Instead of interest being a payment for 'waiting' as in the loanable funds doctrine, Keynes considered it to be a payment a person received for shifting their wealth from liquid money holdings (cash) to some less liquid financial asset.

The concept of **liquidity preference** was introduced by Keynes to explain how people make the choice concerning the composition of their wealth. Holding wealth in its most liquid form (money) has advantages, it is risk free, but it also earns no interest and inflation could deflate its value.

The alternative would be to hold some or all of one's wealth in say, bonds, which attract an interest premium. What might motivate a person to invest in bonds, which are less liquid and carry risk that their capital value could fall with adverse market movements (an excess supply)? The interest rate is the reward for the inconvenience of storing one's wealth in a less liquid form and for the risk of capital loss.

The important point is that for Keynes, the rate of interest does not determine aggregate savings but influences the way that any savings are allocated between liquid and less liquid financial assets.

The interest rate also ensures that the demand for liquidity (money holdings) is equal to the supply of money. We discussed this issue in Chapter 10 *Money and Banking*. As we will see in Chapter 13, the rate of interest also influences the level of national income because it is one of the variables that influence the decision by firms to invest in productive capacity.

The important point of Keynes' attack on the Classical theory of interest was that it debunked Say's Law. Therefore, there was no automatic, market mechanism in a capitalist monetary economy that would ensure that the level of aggregate demand would be consistent with the aggregate supply of goods and services associated with full employment.

It was then understandable that such economies could overproduce goods and services due to the level of aggregate demand expected by business firms being overly optimistic. As a result, firms would have unsold inventories. Their adjustments to the lower realised aggregate spending would lead to cuts in output, income and employment. We formalise these insights over the next three chapters.

Conclusion

In this chapter, we have presented the arguments that John Maynard Keynes used to demonstrate the terminal flaws in the Classical theory of employment and output. His critique involved several steps and we have considered them in turn. Keynes rejected the Classical labour market analysis, which presumes that labour supply and labour demand are independent of one another and independent of the level of aggregate demand. This leads to the conclusion that there are no automatic dynamics operating in the labour market that would ensure movement to full employment. He rejected the Classical notion of Say's Law – the claim that 'supply creates its own demand' – because it presumes that all income received is spent. And, finally, he rejected the Classical argument that the equality of saving and investment is ensured by movement of the rate of interest.

Each of these theoretical errors is related to the Classical belief that the economy moves automatically toward a position of full employment. By exposing the logical flaws in the Classical approach, Keynes opened the possibility that persistent involuntary unemployment is possible due to problems of what he called *effective demand*. In Chapter 13, we introduce Keynes' Theory of Effective Demand, the centrepiece of his greatest contribution, *The General Theory*.

References

Bracha, A., Gneezy, U. and Loewenstein, G. (2015) "Relative Pay and Labour Supply", *Journal of Labor Economics*, 33(2), 297–315.

Dunlop, J. (1938) "The Movement of Real and Money Wage Rates", *The Economic Journal*, 48(191), September, 413–34. Available at: http://www.jstor.org/stable/2225435?seq=1#page_scan_tab_contents, accessed 7 February 2017.

Hansard (1929) HC Deb 15 April, Vol. 227, cc53–6, Winston Churchill 1929 budget speech, delivered Monday, April 25, 1929. Available at: http://hansard.millbanksystems.com/commons/1929/apr/15/disposal-of-surplus, accessed 7 February 2017.

Hicks, J.R. (1937) "Mr Keynes and the 'Classics'", *Econometrica*, 5(2), 147–159.

Keynes, J.M. (1936) *The General Theory of Employment, Interest, and Money*, London: Macmillan, 1957 Reprint.

Keynes, J.M. (1939) "Relative Movements of Real Wages and Output", *The Economic Journal*, 49(193), March, 34–51. Available at: http://www.jstor.org/stable/2225182?seq=1#page_scan_tab_contents, accessed 7 February 2017.

Pigou, A. (1933) *The Theory of Unemployment*, London: Macmillan.

Tarshis, L. (1939) "Changes in Real and Money Wages", *The Economic Journal*, 49(193), March, 150–4. Available at: http://www.jstor.org/stable/2225216?seq=1#page_scan_tab_contents, accessed 7 February 2017.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

13 THE THEORY OF EFFECTIVE DEMAND

Chapter Outline

Learning Objectives

- Understand why employment and output depend on expected effective demand.
- Distinguish between the two components of aggregate demand and the role played by each in determining the point of effective demand.
- Understand how Keynes used the theory of effective demand to expose fallacies in the classical approach to the determination of aggregate employment.
- Acknowledge the policy implications of the theory of effective demand.

13.1 Introduction

In the previous two chapters, we examined the oft-repeated claim among economists that market forces push the economy toward an equilibrium at which supply equals demand. This view is often supported by the common-sense argument that so long as prices are flexible, then they will adjust to an equilibrium level at which the quantity offered for sale just matches the quantity demanded. That equilibrium price is said to be the price that clears the market.

We saw in Chapter 12 that while this type of economic theorising remains influential, it is deeply flawed. It turns out that the simple analogy based on a weekend farmers' market (as made in Chapter 1) and the guidance of an invisible hand cannot be simply scaled up to an economy as a whole in which there are numerous interdependent markets.[1]

The proposition that under flexible prices all markets clear was reinforced by Say's Law that supply creates its own demand. The argument that all income generated from production did not need to be spent in the same period was rejected by Classical economists who invoked loanable funds theory. In turn, Keynes challenged the explicit assumption of loanable funds theory that households' decision to save signalled to firms the intention to engage in higher consumption at some time in the future, which justified firms borrowing funds for investment to provide the capacity to meet this anticipated future demand.

For an economy to maintain full employment when there is a shift in consumption like this, firms would have to know exactly when the future consumption spending would occur and the configuration of goods and services that this future consumption spending might take. If they had that information, then the resources released from the production of consumption goods now could be redeployed to the production of investment goods for future production of consumption goods and services.

However, households save in the form of holding money and other liquid assets which in practice is revealed to firms only when they notice that they have unsold inventory, if they hadn't anticipated this saving. They might respond by reducing production, laying off workers and delaying new investment plans. The economy overall then enters a slump.

In response to Classical theory, Keynes argued that it is the expectation of sales that motivates current production and income generation. Thus, capitalists hire the number of workers they think they will need to produce the amount of output they think they can sell. For Keynes, this is an alternative and better definition of equilibrium.

Rather than arguing that market forces push toward 'market clearing', Keynes said they instead settle at an equilibrium in which firms produce what they think they can sell at a profit. The difference might seem subtle, but it is important. In Keynes' view, the existence of idle resources, including in particular unemployed labour, will not be resolved through the price system. Even if unemployed workers could bid down wages, this would not increase employment if employers do not believe they can sell more output.

Recognising that falling wages in the face of unemployment would reduce the income that could be spent on consumption, Keynes argued that flexible wages might only make matters worse. Assume that employers are pessimistic about sales, so they reduce the scale of production. Also assume that as unemployment rises, wages are bid down in the way conceived by the classical system, as outlined in Chapter 11.

Reduced employment and lower wages (and perhaps reduced working hours for those who do not lose their jobs) combine to lower the wage income of households. With reduced income, workers cut back on consumption. Firms respond with further lay-offs and further reductions in wages and hours. You can see how a vicious cycle might result, with wage and price flexibility only fuelling the downturn. For evidence one only needs to look at the dynamics of the Great Depression in the 1930s when falling wages were associated with falling employment.

Keynes did not rely solely on such evidence in support of his argument. He developed an alternative framework for analysis of the macroeconomic determination of the economy's equilibrium position. In doing so, he abandoned the 'supply equals demand' approach of most economists, and instead, developed a model of effective demand, known as the D-Z approach.

The theory of effective demand, which shows that the macroeconomy can be in equilibrium at less than full employment, is the centrepiece of Keynes' demolition of the Classical macroeconomics model and is as applicable today as it was in 1936 when he published the *General Theory*.

13.2 The D-Z Approach to Effective Demand

The main problem with the market-based approach to equilibrium is that it is fundamentally a 'micro'-based theory and fails when it is applied to the economy overall. When we draw a supply curve for 'widgets' (that is, any produced good) we presume that it is independent of the demand curve for widgets, as well as of demand and supply curves associated with other markets. This **independence** of the two sides of the market is a basic assumption of orthodox microeconomics and is essential for markets to clear when prices adjust to excess demand or excess supply.

Let's go back to the farmers' market again and think about bananas. If we are analysing the market for bananas on a Sunday at 16:59, with a closing time of 17:00, these assumptions might be reasonable.

With one minute left to 'clear' the market, we could imagine that suppliers would offer lower prices, and demanders might raise bids to the point where 'demand meets supply'. After all, unsold bananas would spoil quickly and sellers would be stuck with them if they didn't lower prices, while those who wanted bananas for dessert that evening would miss out if they didn't offer a higher bid. These compatible objectives ensure a convergence at some 'market clearing' price in the final minute of the market day.

We could safely ignore any possible effects of demand on the supply curve. High or low demand would not affect the quantity of bananas at the market because the available banana supply has already been set. The same is true at the fish stand next door. The supply is fixed for the day. It is too late for the banana growers to go back to their farms to harvest more fruit and too late for the fishers to either launch more fishing boats, or to cut back on fishing. Further, we can ignore impacts on other markets. If the equilibrium price of fish falls, lowering the incomes of fishers, this is not going to significantly affect their demand for products at the market (since they've already largely completed their own purchases). And finally, it is too late for suppliers of bananas to switch careers to become fishermen to take advantage of that weekend's brisker sales of fish than of bananas.

However in general, we cannot jump to such conclusions if we are talking about the operation of the economy as a whole over longer periods of time. Here we have an example of the **fallacy of composition**. If over the course of a year the supply of bananas exceeds the demand, falling prices for bananas will have wider effects: the income of banana growers falls, lowering their own consumption. Some switch professions, perhaps pushing wages lower in those jobs. If the low demand for bananas results in a switch to say mango production, jobs and profits may be more plentiful in that sector. On the other hand, if the reduction of banana consumption does not show up as increased consumption elsewhere, the economy as a whole suffers from lower demand and employment.

Keynes realised that if we are analysing the economy as a whole, we cannot simply aggregate up from individual markets, presuming that each is independent. Further, he understood that we cannot presume that the aggregate supply curve is independent from the aggregate demand curve. Firms will only supply output for which they think there is a demand.

For that reason, Keynes needed a completely new framework for analysis. He proposed to use D (aggregate demand) and Z (aggregate supply) curves. These are not the usual demand and supply curves that you see in microeconomics textbooks, which plot quantity against price. That would not be appropriate for aggregate demand and supply, which includes a wide variety of heterogeneous output sold in different markets. We cannot simply add up quantities across a variety of types of goods and services to get a Q that stands for quantity.

Instead, Keynes proposed to use variables appropriate to aggregate analysis. He argued in Chapter 4 of his *General Theory* that there are only two units of measurement which are consistent with macroeconomic analysis: money units and hours of labour input. We can for instance, measure output in terms of either the total money value, or in terms of the number of hours of labour required to produce it.

The first is quite straightforward, and is the way that we measure GDP today. Whether we are summing up the value of the output of computers, automobiles, lumber, or haircuts, we can use the currency value of the final output in each sector. (As we discussed in Chapter 4, comparing levels of GDP over time is more challenging since prices change, but here we are only measuring the nominal value of goods sold over the relevant period.)

Certainly, using labour hours is not as simple. An hour of skilled labour required to produce computer software is not the same as an hour of labour required to sweep the floors of the software firm. However, we can apply weights to labour hours to adjust for skills and education embodied in different types of work performed. While this is not easy, official statisticians actually do conduct calculations similar to this. In any event, in principle we can weight labour hours, with an hour of skilled labour equal to two hours of work requiring lower skill levels (for example), even if in practice this is problematic. This would allow us to measure GDP in terms of weighted labour hours. A greater quantity of GDP would require a greater quantity of labour hours weighted by skill.

For his exposition, Keynes used both the money value of spending and output, and labour hours of input. Instead of price, Keynes used total 'expected proceeds' across all firms. Total proceeds are measured in money terms and are expressed as a function of employment (see Figure 13.1).

Instead of plotting quantity, Keynes used employment, or the total number of labour hours employed. If we assume that labour is weighted by skill, and that each worker works a standard number of hours, then we can use the (skill weighted) number of workers rather than the number of hours worked as our aggregate measure.

Instead of the usual demand curve (linking price and quantity demanded), Keynes proposed the D curve that links expected proceeds to employment. Hence, D represents "the proceeds expected from the employment of N men" (Keynes, 1936: 25).[2]

Obviously, it normally has a positive slope, because as employment rises, firms in the aggregate will expect greater revenues from sales. Even if this might not be true for some firms, it will likely be true of all firms taken together. When employment is strong, expectations of sales revenues will be higher.

Instead of the usual supply curve (linking price and quantity produced), Keynes substituted the Z curve, again linking expected proceeds to employment: Z "is the expectation of proceeds which will just make it worth the while of the entrepreneurs to give that employment" (Keynes, 1936: 24).

This also slopes up because firms require higher expected revenue to induce them to hire more workers.

What we need to keep in mind is that Keynes is proposing to replace 'market clearing equilibrium' with his notion of 'effective demand'. Firms will employ the number of workers that they think they will need to produce the amount of output that they expect to sell at a profit. They will hire more labour and produce more output if they think the sales revenue will be higher. This leads to a positive slope for the Z curve. The slope of the Z curve depends on the conditions of supply, and the rising slope indicates that as employment increases, firms require that income rises faster than their employment costs. Firms can face resource bottlenecks as they expand output.

Beyond some point, as firms increase production, they find it more expensive to hire labour with just the right set of skills. They might have to pay more to purchase larger quantities of other crucial inputs from suppliers who face their own bottlenecks. When firms form the expectations that influence the amount of labour they choose to employ, they consider the conditions they will face in hiring labour and purchasing the other means of production. When aggregate employment is high and labour markets are tight, they require higher expected proceeds. We thus expect the slope of the Z curve to increase as we move away from the origin due to perceived tightness and expected rising costs of increasing output.

As we will discuss later in this section, with more labour hired, there will be more wages generated, which might actually generate more sales revenue. It is this expectation that gives the D curve a positive slope. However, here we need to emphasise that the D and Z curves concern expectations that influence the determination of employment at a point in time. Firms cannot know at the time they make their decision what the actual outcome will be. They might find that they were overly optimistic (profits are too low or non-existent), or overly pessimistic (they could have sold more and reaped greater profits if they had employed more workers), or perhaps just right. The outcome, in turn, can affect their expectations in the future and hence future employment decisions.

Finally, we need to understand that the D and Z curves are for the economy as a whole, and do not represent demand and supply curves for an individual market. Effectively what we are doing is summing across the employment decisions of all the firms in the economy to find the aggregate level of employment. Each firm hires up to the point where it employs the number of workers it needs to produce the amount of output it expects to sell at profit, taking into account expected revenue and expected costs at each level of employment. In Chapters 15 and 16 we will return to a more detailed discussion of the analysis of aggregate demand and supply.

Figure 13.1 depicts the D-Z curves. The curves are represented as non-linear functions, and we will discuss why they might be nonlinear. They are drawn for a given money wage level, W_0. The intersection of the two curves is labelled point A. This is the point of effective demand, which is the aggregate level of employment consistent with the expectations that lie behind the employment decisions of firms.

To the left of point A, the D curve lies above the Z curve. This means that at any point to the left of A the expected proceeds from hiring the corresponding amount of employment is greater (shown on the D curve) than the cost required by firms to hire that amount of employment (shown on the Z curve). In other words, D > Z. Rational firms expect to increase revenues and profits by employing more workers.

However to the right of point A, the expected proceeds from hiring more workers are less than is the outlay required to induce firms to actually hire more since the Z curve lies above the D curve (Z > D). Rational firms will not employ more workers than they think they need to produce the amount of output they expect to sell profitably.

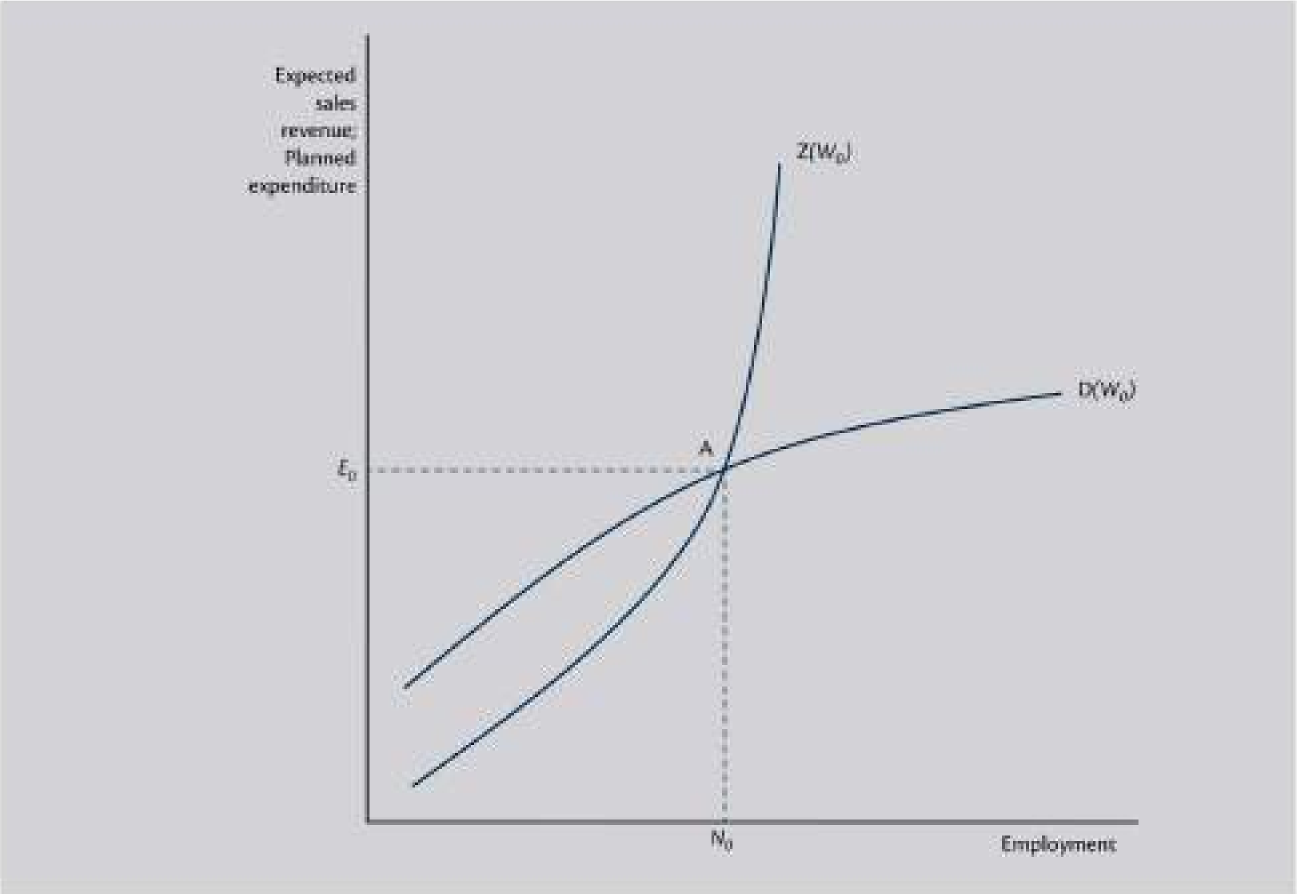
We cannot say for sure that all points to the right of point A actually result in losses to firms. All we can say is that the expected proceeds are below what they want to obtain from employing the corresponding amount of labour.

The equilibrium is at point A: firms are employing the amount of labour that is expected to produce the expected revenues that are required to induce them to employ that amount of labour. We do not have to call this profit maximisation. Nor is it a market clearing point of equilibrium. It is simply the equilibrium level of employment – the number of workers at the aggregate level that firms wish to employ given their expectations of sales. If they became more optimistic about sales, they would hire more; if they were more pessimistic they would hire fewer. But given their expectations, point A is just right.

Keynes called A the point of **effective demand**. By that he meant that employing N_0 workers (shown on the horizontal axis) is expected to produce the corresponding level of proceeds (E_0) (shown on the vertical axis). Higher employment would be expected to produce more income and hence higher 'demand'. However, actual sales are expected to be below what firms require to employ more workers (Z > D). Lower employment would be less desirable because expected revenues exceed required revenues (D > Z), providing an incentive to increase employment.

Only at point A do we have an equilibrium level of effective demand. Macroeconomic equilibrium thus occurs for the current given money wage rate (W_0) at N_0, which determines total employment in the economy.

Figure 13.1 Keynes' D-Z aggregate framework



13.3 Introducing Two Components of Aggregate Demand: D1 and D2

Keynes divided aggregate demand D between two categories: D1 and D2. D1 rises with employment (and income). This is normally associated with consumption, which is assumed to be a function of income and thus employment. On the other hand, D2 covers spending that is 'autonomous' to employment and income. That means that the level of D2 does not tend to change when employment and income change.

Thus, the determinants of D2 are variables other than employment and income. Keynes included investment spending in this category. The reason is this: investment is mostly a function of expected future profits, so it depends on what Keynes called 'the state of long-term expectations'. While these expectations may be affected to some degree by the current state of the economy, profits that are expected many months and years into the future will be the main influence on these long-term expectations.

Consumption is determined mostly by current disposable income and the propensity to consume out of income. Further, the propensity to consume out of income is normally thought to be greater than zero but less than one. This means that if income rises by a dollar, consumption changes by some positive amount, but by less than a dollar. As Keynes put it, "the psychology of the community is such that when aggregate real income is increased aggregate consumption is increased but not by so much as income" (Keynes, 1936: 27). Keynes goes on to argue that the marginal propensity to consume falls as income rises (he says that "the richer the community" the lower the propensity to consume, which he labels "the fate of Midas") (Keynes, 1936: 31; 219).

We can presume that as employment rises, consumption rises because higher employment generates more income. If there is a fairly stable but slowly declining relation between an increase of employment and the rise of income generated by that employment, then the propensity to consume associated with a change in employment will be the product of the rise in income and the propensity to consume out of that income. This means that the D1 curve will have a positive but declining slope like the D curve in Figure 13.1: beyond some point the slope will decline as the propensity to consume tends to fall as employment and income rise. This is often called a **demand gap**: hiring more workers and paying them wages increases aggregate demand, but by less than the income that is paid out by employers.

Note that all the wage income generated by employing workers represents a cost to the employers who employ those workers. In the aggregate, employers cannot recover their additional wage expenses if newly employed workers only spend a fraction of the wages they are paid. The argument applies more generally to all incomes that are generated by an expansion of production, wages, profits, or rents if the propensity to spend out of that new income is less than one. In other words, not all of the additional income generated by higher production will be spent on consumption.

These costs of production, that is the income that is paid in the form of wages, profits, and rents during the production process, underpin the Z curve since they are costs that need to be recovered to induce producers to employ workers to produce a particular level of output. There would be no equilibrium point of effective demand if there was only spending on consumption (*D1*) because expected costs (the main determinant of the proceeds required to induce entrepreneurs to employ workers) would exceed expected revenues from employment (that is, from the sale of output to satisfy *D1*) So long as the propensity to consume out of income is less than one. **This is the demand gap.**

What about D2? Since by assumption, D2 is not a function of employment or income, if we were to graph it in Figure 13.1 it would be depicted as a horizontal line. D2 is associated with investment in the means of production, but it can include government purchases as well as export demand (which might depend on employment abroad, but is unlikely to be associated with domestic employment).

We add D2 to D1 to obtain the total aggregate demand, shown as the D curve in Figure 13.1. D2 can fill the 'demand gap', as long as it is positive.

We have already said that the point of effective demand occurs where the D curve meets the Z curve. Given expectations it is the only possible point: firms employ the number of workers they think they need to produce the output they think they can profitably sell.

In Figure 13.1, at point A we find that D1 plus D2 is equal to Z. It is at this one point – the point of effective demand – that the aggregate expected proceeds (D = D1 + D2) equal the proceeds, Z, that are required to induce firms to hire just that amount of labour. Keynes referred to this model as the **General Theory of Employment**.

In Chapter 15 we will present a more detailed discussion of the components of aggregate demand, and will frame the exposition in terms of aggregate income rather than aggregate employment. This is the method usually adopted by economists today and is more consistent with the national income and product account methodology we discussed in Chapter 4. However, Keynes' goal in developing the D-Z framework was to show how the equilibrium level of employment is determined without reference to a labour market in which flexible wages are claimed to function to 'clear' the market. Keynes emphasised that the point of effective demand need not correspond to full employment.

13.4 Advantages of the D-Z Framework

The D-Z framework developed by Keynes has several benefits if we are addressing the economy as a whole. First, employers make production decisions based on expectations of nominal proceeds. Our supply curve, Z, is forward looking. Unlike the supply curve in a market for bananas at 16:59 on a Sunday, which represents past supply decisions and sales earlier in the day, the Z curve represents the expectations of suppliers looking forward in time to the period in which they will undertake production.

The Z curve shows the different levels of revenues (proceeds) that must be generated to induce firms to hire the corresponding number of workers. Firms also hold expectations about the sales and hence revenues that will be forthcoming from different levels of employment, shown on the D curve. These expectations will be mediated by actual macroeconomic outcomes. If firms have underestimated revenues and hired too few workers (producing too little output), then inventories will be run down. For example, retail stores will find that they cannot keep sufficient stock on hand to meet customer demand, so the shelves begin to empty. The stores place new orders with suppliers, who in turn place orders with manufacturers. If manufacturers believe the high demand will be sustained, they increase employment and output. In terms of Figure 13.1, the D curve shifts out and the point of effective demand will be to the right of point A.

On the other hand, if employers had overestimated revenues, inventories pile up and retailers reduce orders. Producers find they are producing too much and might reduce employment if they think the sales slump will continue. The point of effective demand shifts left in Figure 13.1.

We emphasise that the D-Z framework is a tool that we as economists can use to think about the determination of employment at the aggregate level of the economy. Individual firms do not use the framework to determine their own level of employment. Except for the biggest firms, the impacts of a firm's hiring decisions on the aggregate level of employment and spending can be ignored. While large firms certainly do have departments that forecast aggregate demand for use in planning production of their range of products, smaller firms focus more on their own individual situation.

As macroeconomists, we are principally concerned with the determination of total employment and output. We also recognise that the employment decisions of individual firms impact on the sales experienced by other firms.

Thus, if the sum of employment decisions made by individual firms leads to more employment, then aggregate sales would be expected to be higher as production decisions are modified over the course of a year. Using the D-Z analysis allows us to take into account the interdependence of the demands and supplies, as well as the impacts of production decisions of one market on others.

The macroeconomic demand for labour

The D-Z framework enables the derivation of the macroeconomic demand for labour schedule as a function of the money wage. A complete treatment can be found in Chapter 14, and we just provide some introductory comments here.

The money wage impacts both the supply and demand sides of the economy. A change in the money wage changes unit labour costs at the firm level, which in turn, affects the industry supply curve and thus shifts the aggregate Z function.

At each employment level, a rise in the money wage will push the Z function up because the firm now requires more sales revenue to sustain the same employment and output levels to satisfy their profit expectations.

But, a higher money wage also means that at each employment level, the nominal incomes paid out are higher and this stimulates nominal aggregate demand (in the first instance via higher aggregate consumption). Therefore, the D function shifts up when the money wage rises because firms will expect higher sales revenues (and shifts down if the money wage is cut in expectation of lower sales revenues).

The point of effective demand will thus be sensitive to the magnitude of the money wage. This in turn sets the level of employment that firms will offer and the level of output and national income generated by the economy.

We can draw a family of aggregate D and Z functions (and points of effective demand) corresponding to different money wage levels. This allows us to relate each money wage level with a particular point of effective demand *and* aggregate employment. The importance of the recognition that changing wages causes both the D and Z curves to shift is what caused Keynes to reject the Classical theory of employment, which considered the marginal productivity function to define the macroeconomic labour demand function without reference to the goods and services market.

The approach adopted by Keynes and those who followed him situated the derivation of the macroeconomic labour demand function in the goods and services market, by examining how money wage movements influence the point of effective demand. Thus, when considering a change of money wages we must take account not only of the impact on the supply side, but also of the impact on the demand side of the goods and services market.

What this means is that we must reject the simple 'classical' labour market approach which claims that employment is determined by an upward sloping labour supply curve, downward sloping labour demand curve, and a flexible real wage that ensures full employment. That approach is inconsistent with the recognition that the outcome in the labour market is dependent on firms' expectations about the proceeds to be generated from employment, that is, aggregate spending.

13.5 The Role of Saving and Liquidity Preference

In Classical theory, it was assumed that when one chooses not to spend one's income on current consumption (that is, to save income), that is matched by a decision to spend on investment (which will produce output to be consumed in the future). This is why Say's Law holds: since a decision to save is matched by a decision to invest, there cannot therefore be a demand gap. All income always gets spent, either on consumption goods or investment goods and we can say that 'supply creates its own demand'.

Keynes argued that this relationship is false. A decision to save does not automatically generate a decision to invest. It might be true that an individual's decision to save is a decision to postpone consumption. However, that individual can postpone consumption by saving in the form of money or financial assets which are liquid, rather than by placing an order for goods and services to be provided at a particular time in the future. This is why saving typically opens up a demand gap.

Keynes argued that in the real world there is a preference for liquidity, a desire to accumulate holdings of liquid assets such as money and financial assets. This preserves options, since an individual who holds money can decide later when, and where, and what to consume. If one is uncertain about the future, it makes sense to save in a liquid form.

According to Keynes, this makes a big difference to our analysis of the aggregate economy. Uncertainty lies behind both the D and the Z curves. Firms form their expectations about demand for their output while realising that their best guesses can turn out to be wrong. If their uncertainty is too great, they have the option of forgoing hiring and curtailing production.

Most importantly, if they are pessimistic about the future, they will decide to hold off investing in plant and equipment. Just as the household has the option of holding money and financial assets, the firm has the option of purchasing financial assets rather than investment goods. That will impact on the D2 curve because spending on capital goods generates employment and effective demand. On the other hand, the purchase of financial assets might create a few jobs in the financial services sector, but this will not offset all the jobs lost in the investment goods sector.

For these reasons, a desire to postpone consumption can reduce D1 without increasing D2, which creates a demand gap. Thus, labour displaced from producing goods to satisfy D1 does not find jobs producing the output included in D2.

13.6 The Demand Gap Arguments and Policy Implications

From the previous sections, we understand that expected D1 spending at any given level of employment will generally fall short of what is necessary to draw forth the supply (shown on the Z curve) corresponding to that level of employment. This demand gap needs to be filled by D2 spending to produce a point of effective demand at that level of employment. So far, we have restricted our analysis to investment spending as a component of D2 spending.

We have also noted that because investment spending depends on the expectations of firms concerning future profits to be made from additional plant and equipment, there is no assurance that investment will be high enough to fill a particular demand gap. However there are two other kinds of spending that we need to consider: government spending and net exports. These can also reduce the demand gap.

Let us first deal with net exports. As we will discuss in Chapter 15, imports represent a leakage out of the domestic economy. If consumers spend income on foreign-produced goods and services, this reduces domestic demand. A portion of imports could be considered to be somewhat independent of domestic income while a significant portion is considered to be a function of domestic income. This means that imports would affect both the D1 and the D2 curves, which would shift the D curve down in Figure 13.1, adding to the demand gap. As domestic income rises, the demand for imports rises. This propensity to import might even increase as income rises since more imported luxury consumption goods might be demanded.

On the other hand, exports add to domestic demand, because foreigners purchase goods and services produced domestically. Export demand is relatively independent of domestic income. Hence, exports show up primarily in the D2 curve. Net exports are defined as exports less imports. If positive, net exports mean that the D2 will be higher, helping to close a demand gap. All else being equal, this means the point of effective demand in Figure 13.1 will be further out to the right, with a higher level of employment.

This is why nations often pursue a strategy of promoting net exports. By raising aggregate demand through exports, the level of effective demand is higher and that generates more employment.

Obviously, it is impossible for all nations to actually run positive net exports at the same time. It is a zero-sum game since globally exports must equal imports. The strategy of promoting net exports to close demand gaps is called mercantilism. It is a **beggar thy neighbour** policy because those nations that succeed do so at the expense of other nations that end up with trade deficits. As Keynes pointed out, it leads to international hostility, including trade wars, and even to war.

There is an alternative strategy that does not pit nation against nation. Another source of demand is government spending, or fiscal policy more generally. Let us first consider tax policy and then move to government spending.

We have so far ignored taxes, but it is clear that taxes will reduce the spending of the private sector. Taxes primarily hit consumers, and so reduce the D1 portion of spending. Many taxes rise with income. Other taxes are not directly related to household income, and some impact on businesses and investment decisions. In that case, they might lower the D2 type of spending.

Government can reduce the demand gap by lowering taxes. However, this can be a rather blunt instrument because reducing taxes does not directly increase demand. To be effective, the tax cuts need to increase the willingness of households and firms to spend on domestic output. If the tax cuts lead to increased saving, or paying down debt, or higher imports, they do not stimulate domestic demand. Only if the tax cuts induce spending on domestic output will they raise demand by increasing D1 and/or D2.

Increasing government spending can also be used to fill demand gaps. And because government spending can be directed toward domestic demand, it is a powerful policy instrument. Let us consider three alternative approaches.

First, government can purchase domestic output directly from domestic producers. By ordering goods and services, government adds to demand and generates employment in the private sector. This is represented in Figure 13.1 as an upward shift of the D2 curve, pushing the point of effective demand where the D curve crosses the Z curve out to the right. Employment and output are higher.

Second, government can increase transfer payments to households. This could be in the form of higher unemployment compensation, welfare, or social security payments, for example. This will stimulate consumption by households. However, as in the case of a tax cut, there is no guarantee that households will increase their spending on domestic output. The impact on demand for domestic output will depend on what those who receive the transfer payments decide to do with them.

Some of the increased transfer payments will leak out of the economy through higher saving and more imports (and higher taxes). Still, it is very likely that if the transfers are targeted to households with low income, consumption will rise and the D1 curve will shift up, generating more employment and greater domestic production.

Finally, government can directly hire the unemployed. This will directly increase employment and generate more household income. Again, that is likely to generate more household consumption, shifting the D1 curve up. This is the well-known notion behind the government spending multiplier (see Chapter 15 for a formal treatment of the multiplier). There will also be some leakages to imports, saving, and taxes, but the net impact will be to raise employment (directly) and consumption.

We must close this section with a caveat, however. Recall from previous sections that the D curve (as well as its components, D1 and D2) is framed in terms of expectations: firms hire labour based on their expectations of sales. In this section we have discussed various injections and leakages that come from the investment, government, and foreign sectors. These can help to close (or open up) the demand gap created by a propensity to consume which is presumed to be less than one.

We need to remember, however, that what matters for the determination of the point of effective demand is expected revenues relative to costs. If an announced government policy of tax cuts or of spending increases were to raise sales expectations of employers, the point of effective demand would be farther out to the right, indicating a higher level of employment. However, it is possible to conceive of a situation in which the expectations of employers do not react in a positive manner to such an announcement; maybe they do not believe it, or maybe they expect negative consequences (inflation that raises costs more than expected revenues, or monetary policy tightening). In that case, they may not employ more workers (the point of effective demand is not farther to the right).

The point is that we should not view these D and Z curves as conceptual devices that we can shift about in a mechanical fashion to 'fine tune' the economy. Rather, they are analytical devices that help us to understand how the aggregate level of employment and output is determined.

Conclusion

In this chapter, we have examined Keynes' exposition of the theory of effective demand by making use of his D and Z curves. These relate employment to the point of effective demand. Keynes' approach takes account of the fact that employment affects income and output, and hence affects the demand for workers. Generally, greater employment leads to greater sales. The higher the point of effective demand, the greater the level of employment and the greater the expected sales.

Keynes' D-Z framework provides us with an alternative to the Classical labour market analysis. Rather than determining the aggregate level of employment in the labour market through independent labour demand and supply curves (as Classical theory does), Keynes argued that aggregate employment is determined at any given time by the point of effective demand.

The D and Z curve exposition also provides advantages over the typical aggregate demand and aggregate supply curves, which relate output to an aggregate price level. First, the traditional aggregate demand-supply model aggregates up from the micro level, presuming independence of the two curves. While the demand-supply analysis might be fine for a local market for bananas late on a Sunday afternoon, it is not appropriate to assume that supply and demand are independent for the economy as a whole.

Since the supply of output affects employment and income, it must affect demand. Keynes took this interdependence into account by positing that the point of effective demand depends on the level of sales expected as a function of the level of employment. Firms then hire the amount of labour they need to produce the amount of output they think they can sell.

Keynes rejects Say's Law, the proposition that supply creates its own demand, because the income that is generated in the production process ('supply') does not have to be spent ('demand'). One can save part of one's income in a liquid form – financial assets – rather than purchasing output.

This opens the possibility of a demand gap at any given level of employment, that is, supply is greater than demand. Keynes' method allows us to see how it is possible for aggregate spending to be equal to aggregate income (which is necessarily true at the aggregate level) without invoking Say's Law.

At the point of effective demand, the income that is generated by production (supply) is expected to equal the sales that will result from employing workers to produce it (demand).

However, Keynes insisted that the point of effective demand need not be consistent with full employment. Indeed, the point of effective demand could be consistent with any level of employment.

One must not forget the important point that Keynes' D-Z framework is underpinned by expectations. The point of effective demand (point A in Figure 13.1) is an equilibrium where the level of employment is consistent with the level of aggregate proceeds that satisfy two conditions: these are the proceeds that are expected to be generated from precisely that level of employment, and they are the proceeds that are required by firms to employ precisely that number of workers.

However, it would be a mistake to jump to the conclusion that the level of employment consistent with the point of effective demand (N_0 and point A in Figure 13.1) will actually generate the level of sales expected. In other words, our diagram is *ex ante*, showing what is expected, rather than *ex post*, showing what actually happens with respect to sales. It does show how much employment will be offered by producers in the aggregate. Sales might turn out to be higher than expected (great!) or lower than expected (disappointing!). That outcome might then cause firms to change their expectations, leading to a different point of effective demand in the next period. The level of employment consistent with their new expectations could be higher or lower than the equilibrium N_0.

Hence the point of effective demand (and thus employment) will likely change as we move through time, depending in part on whether expectations turn out to be validated. Expectations can also be changed by policy pronouncements by government; a new spending programme, or changes to taxes, for example. The opening of new markets abroad, new technological innovations, discoveries of new natural resources, and other such events can also change the state of expectations and hence the point of effective demand.

Reference

Keynes, J.M. (1936) *The General Theory of Employment, Interest, and Money*, London: Macmillan, 1957 Reprint.

Endnotes

1. The formal demonstration that a simultaneous general equilibrium could exist in a market economy turned out to be quite elusive, requiring complex mathematics and very limiting assumptions (including the assumption that there is no money). Simply stated, it cannot be demonstrated for a real world economy.
2. Here we are presuming that the number of workers employed can stand in for the number of hours worked, adjusted by skill level. This requires the use of an average number of hours worked per day, and weighting of skill level to create an average measure of productivity per hour worked. Note also that the use of the term 'men' for workers was a sign of the social attitudes that pervaded in his day.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

14 THE MACROECONOMIC DEMAND FOR LABOUR

Chapter Outline

Learning Objectives

- Understand the derivation of the macroeconomic demand for labour.
- Explain why unemployment is the outcome of insufficient effective demand.
- Acknowledge that cuts to money wages are unlikely to address persistent unemployment.

14.1 Introduction

In this chapter we build on Keynes' derivation of the point of effective demand discussed in Chapter 13 to derive formally the macroeconomic demand for labour by considering the impact of changes to the money wage on the aggregate demand (D) and aggregate supply (Z). This analysis generates a schedule relating employment to the level of the money wage and replaces the flawed Classical marginal productivity theory, which fails to recognise the interdependency of aggregate supply and demand schedules.

The macroeconomic demand is then juxtaposed against a schedule depicting the macroeconomic supply of labour.

14.2 The Macroeconomic Demand for Labour Curve

The interdependency of aggregate supply and demand

> **REMINDER BOX**
>
> Recall that Classical employment theory considered unemployment (beyond frictional levels) to be the result of the real wage being higher than the equilibrium level associated with labour demand equalling labour supply. Classical theorists concluded that flexibility of money wages (and hence real wages) would continuously clear the labour market and maintain full employment.
>
> In the Classical approach, given that profit-maximising firms are assumed to be constrained by diminishing marginal productivity of labour, they will only employ more workers if the real wage falls because each additional worker hired is less productive than the last. Profit maximisation requires that the real wage (the cost of the additional unit of labour) be equated with the marginal product (the contribution to output of the last unit of labour).
>
> When money wages fall, the marginal cost of production falls (assuming productivity is unchanged) and firms' output supply curve shifts out, that is, they are prepared to supply more at each price level because their marginal costs are lower.
>
> Classical theory considered the aggregate supply curve to be the sum of the industry supply curves, which in turn were considered to be aggregates of the firm supply curves. As a result, when money wages fell, the supply curves of firms, the industry and all industries would shift outwards.

In the Classical system, a money wage cut leads to an expansion of output because firms are prepared to hire more workers and to supply more output at the given price level. Keynes rejected this reasoning, which required that the aggregate supply curve (the sum of all the firm supply curves) shifts out as money wages fall and traces out a path along a **fixed** aggregate demand curve which is considered to be independent of the aggregate supply curve. We considered this model in terms of the theory of effective demand outlined in Chapter 13. The basic reason for Keynes' rejection of this argument was that he believed that a cut in money wages would impact negatively on total spending.

Once again this is an example where specific to general reasoning provides the wrong answer. What might apply at the single firm or even industry level does not necessarily apply at the aggregate level. That is the basis of the fallacy of composition, which we explained in the context of the paradox of thrift in Chapter 2.

At the individual firm level, the output supply curve would shift out (meaning it would be prepared to supply more output at each price level) after it cut the money wage of its workers. The firm would not expect a shift in the demand for its product as a result of the money wage cut, but would expect to sell more output due to charging a lower price.

But what if all firms cut money wages? The Classical theory of employment focused on only one aspect of money wages: that they were a cost of production and influenced the supply side of the economy. However, Keynes noted that money wages are also a significant component of a worker's income and by aggregation, national income.

Given that consumption spending is directly tied to national income, and investment spending is also likely to fall if consumption spending fell, Keynes argued that the demand curves (at the firm, industry and aggregate) level would shift inwards (spending at each price level would be lower) after a money wage cut.

While firms might enjoy lower unit costs, they also faced a declining demand for their goods in general, because the lost income resulting from an economy-wide money wage cut would be significant.

This insight suggests that the **output demand and supply curves are interdependent**. This interdependency also negates the Classical construction of the aggregate marginal productivity curve as the macroeconomic demand curve for labour.

In Classical employment theory, a money wage cut (indicating that the labour supply curve has shifted out) leads to a movement along a fixed marginal productivity curve if the price level is constant or falls by less than money wages. However, once we recognise the interdependency between the demand and supply sides, it is clear that the marginal product curve does not represent the demand schedule for labour.

In Chapter 13, we outlined Keynes' model of employment. We showed that the point of nominal effective demand is found at the intersection of the aggregate demand and aggregate supply price curves. At each point on the aggregate supply price curve (Z), the revenue required to induce that amount of employment should be sufficient to cover all production costs and desired profits at the relevant employment level. So moving up a given Z function means the firms will be employing more workers and producing more output and generating more income.

The aggregate demand price (D) function describes how much expected demand there will be in the economy at different employment and (implied) income levels. We assume that the money wage is fixed along any given D function. Thus, as employment rises, total national income rises and expected spending also rises accordingly.

The intersection between aggregate demand and aggregate supply, which defines the point of effective demand, will thus depend on the magnitude of the prevailing money wage. This in turn sets the level of employment that firms will offer and the level of output and national income expected to be generated by the economy. The point of effective demand occurs where all individual firms are employing the amount of labour they think they need to produce the amount of output they expect to sell.

In Section 13.4, we saw that changes in money wages influence the point of effective demand by shifting both the Z and the D curves in Keynes' model of employment which recognised the interdependency of the two schedules. This observation provides the clue to deriving a **macroeconomic labour demand curve**.

Money wage changes and shifts in effective demand

Figure 14.1 illustrates the case proposed by the classical economists. Figure 14.1(a) shows the family of D and Z functions for different money wage levels, W_0, W_1, and W_2, with $W_0 < W_1 < W_2$. Hence we can see the points of effective demand corresponding to different money wages. For example, at money wage W_0, aggregate demand D(W_0) and aggregate supply Z(W_0) jointly determine the point of effective demand which corresponds to employment N_0. In Figure 14.1(b), employment is plotted against money wages, so N_0 represents the macroeconomic labour demand corresponding to money wage level W_0.

As shown in Figure 14.1(a), at each employment level a rise in the money wage (above W_0) will push the Z function up because firms now require more sales revenue to sustain the same employment and output levels to satisfy their profit expectations. But the higher money wages also mean that at each employment level, incomes are higher and this raises the revenue expected at that level (the D curve shifts out). Therefore, the D function shifts up when the money wage rises (and shifts down if money wages are cut).

We can draw a family of aggregate D and Z functions (and identify corresponding points of effective demand) for different money wage levels. This in turn, allows us to relate each money wage level with a particular point of effective demand **and** aggregate employment. Then each point of effective demand yields a money wage and employment combination which is a point on the macroeconomic labour demand function, as shown in Figure 14.1(b).

Under Classical theory the marginal productivity function defines the macroeconomic labour demand function without reference to the goods and services market. In contrast, in the D-Z framework we need to know the point of effective demand corresponding to each money wage to determine the quantity of employment offered.

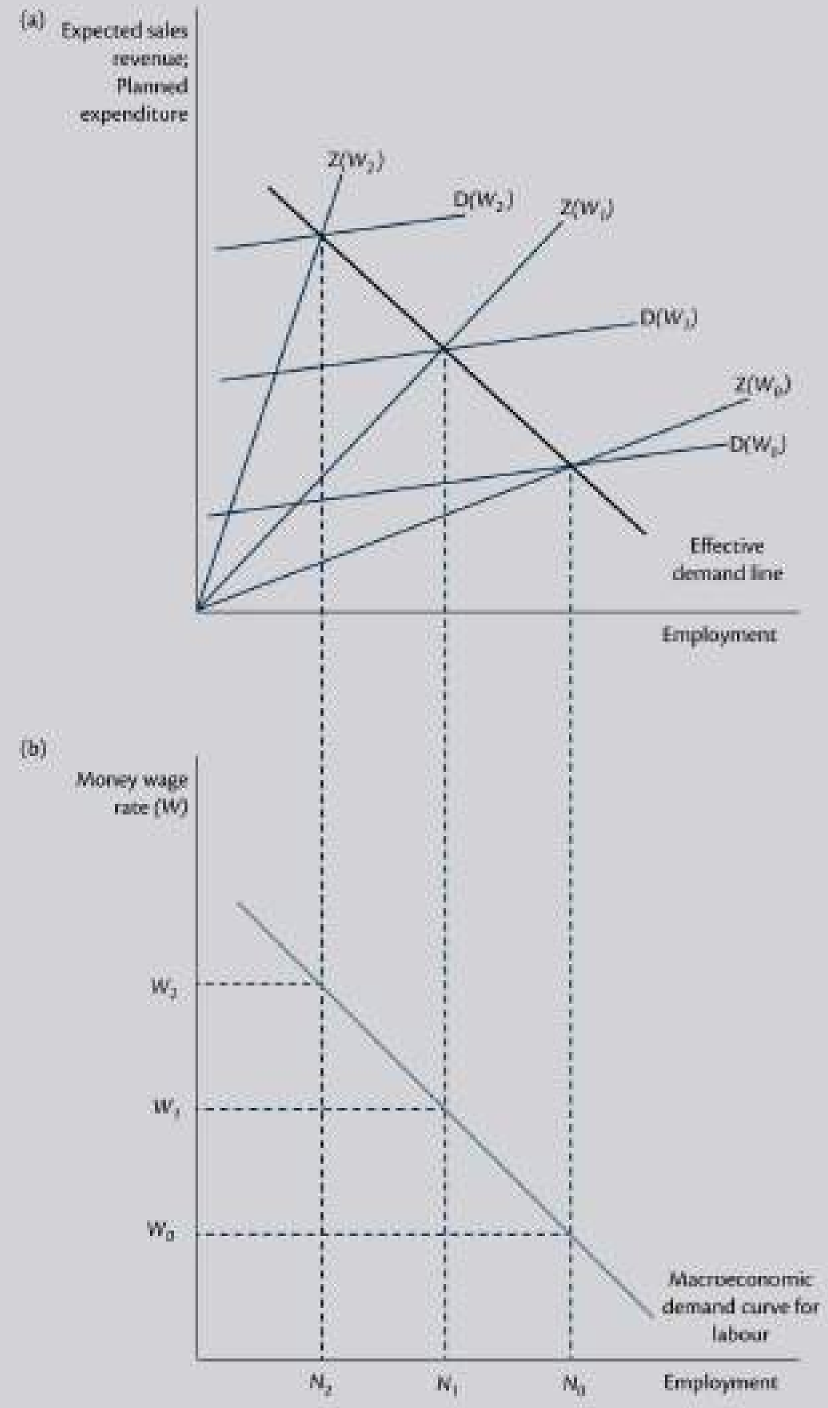
We now need to consider whether there is a particular relationship between wages and employment. If the money wage changes, three different scenarios can be identified:

- The D function shifts by less than the Z function and thus employment falls(rises) when the money wage rises(falls).
- There are equivalent shifts in the D and Z functions and employment does not change.
- The D function shifts by more than the Z function and thus employment rises(falls) when the money wage rises(falls).

In 1956, American economist, Sidney Weintraub, said that the first possibility was consistent with the traditional argument that when money wages rise, employment falls. He called this the 'Classical' case, which reflects the proposition in Classical employment theory that unemployment can be eliminated by cutting money wages.

Figure 14.1 depicts the 'Classical' case, showing the family of D and Z functions for different money wage levels and the corresponding employment rates.

Figure 14.1 The 'Classical' case



The Z function shifts upwards by more than the D function shifts as the money wage rate rises. The points of effective demand, as shown by the effective demand line, yield a downward-sloping **macroeconomic demand for labour curve**, which demonstrates the relationship between money wages and employment.

Keynes attacked the Classical employment theory because he did not believe that employment was particularly sensitive to money wage movements. In terms of the three scenarios for the relative movements in the D and Z functions, Keynes' position corresponds to the second option, namely that the two functions shift equivalently.

Figure 14.2 captures this situation, which Weintraub characterised as the 'Keynesian case'. You will see that under these assumptions, the macroeconomic demand curve for labour is vertical and invariant to movements in money wages.

The third option – that the D function shifts by **more** than the Z function when the money wage changes – was referred to by Weintraub as the 'underconsumptionist' case. Some economists believed that higher money wages would stimulate employment because of the rise in consumption. Figure 14.3 depicts the 'underconsumptionist' case.

Note for simplicity, linear 'curves' are used in Figures 14.1, 14.2 and 14.3.

Figure 14.2 The 'Keynesian' case

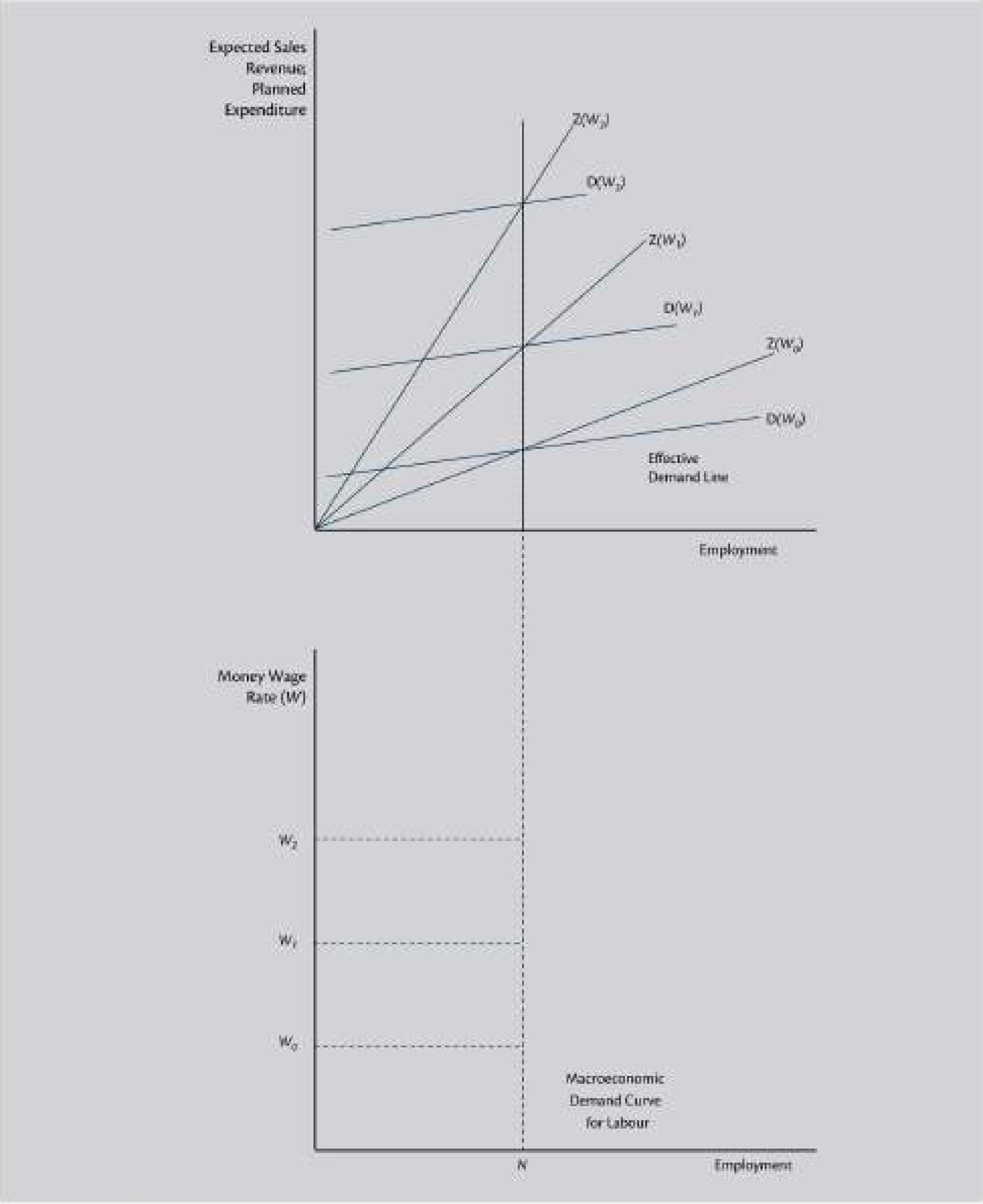
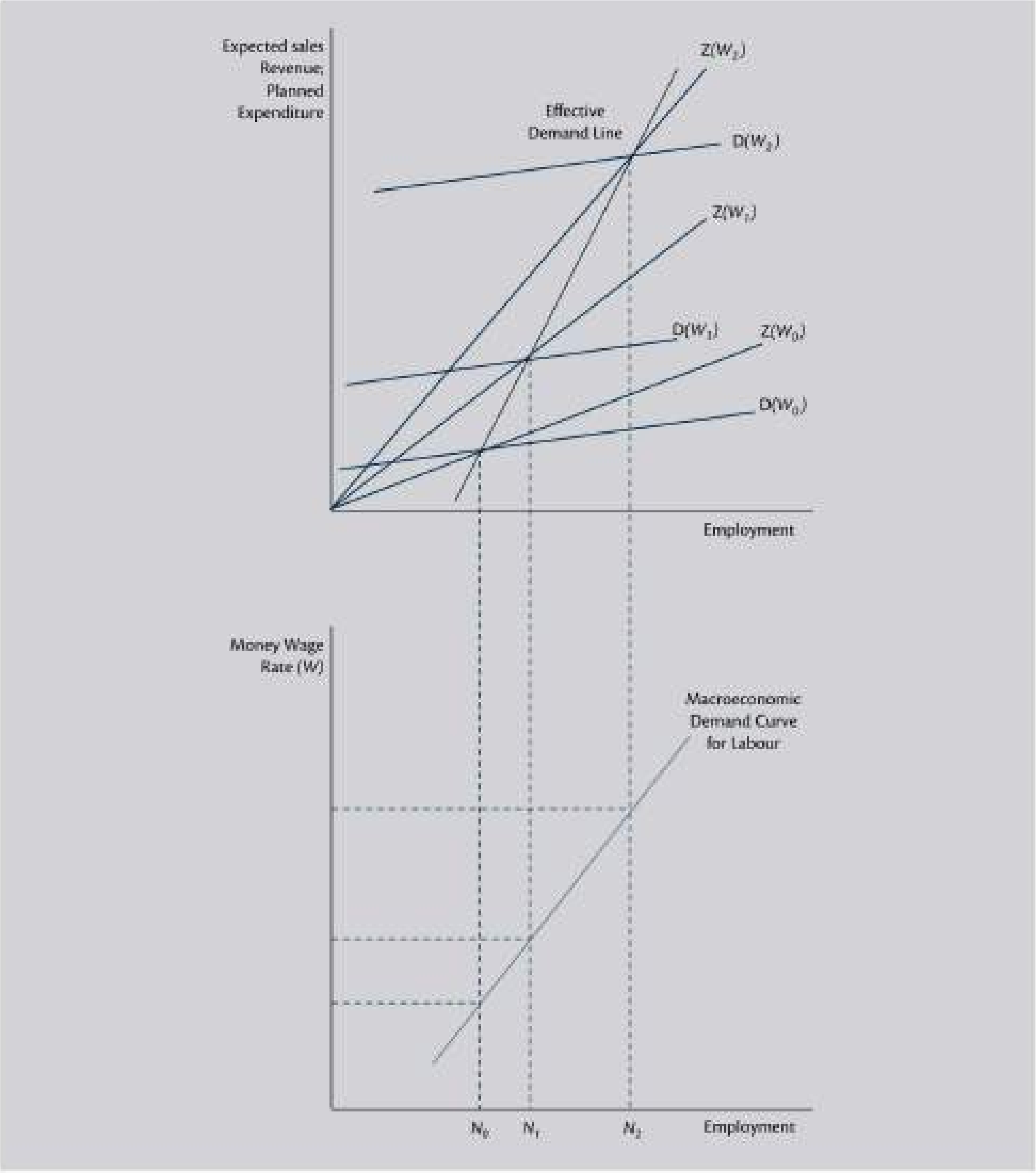


Figure 14.3 The 'underconsumptionist' case

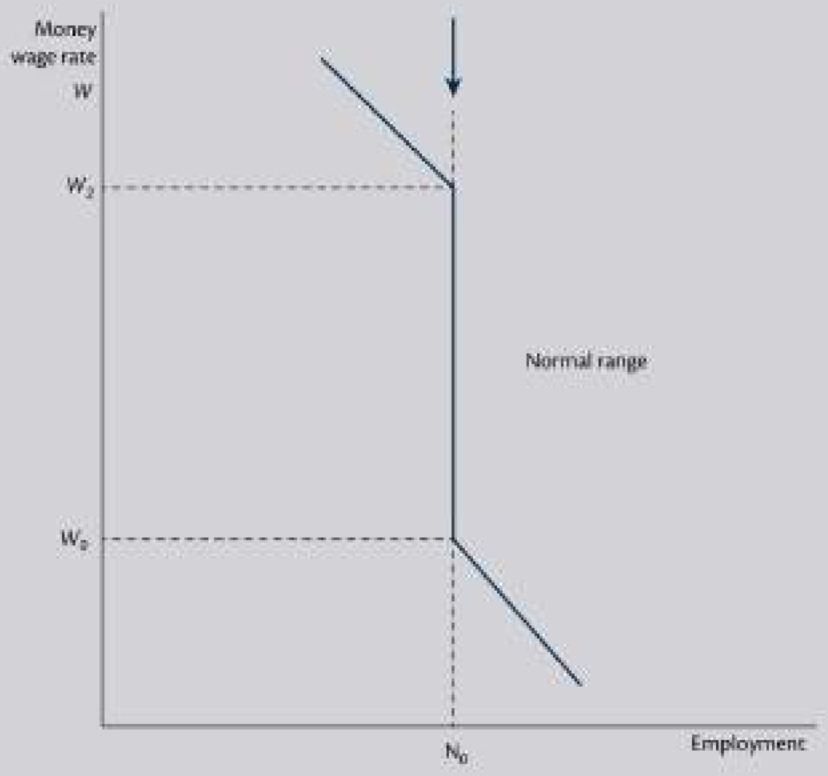


Weintraub (1956: 842) said in relation to these three cases that "from the standpoint of economic policy their implications are vastly different."

If for example, the real world was more like the Keynesian or the underconsumptionist cases, then trying to cure unemployment by cutting money wages would fail, and in the underconsumptionist case, would be a disaster for employment.

Weintraub considered the real world was more like that depicted in Figure 14.4 which is what we call a generalised macroeconomic demand curve for labour with three distinctive segments. The blue arrow signifies that employment is a function of effective demand, which is determined outside of the labour market.

Figure 14.4 A generalised macroeconomic demand curve for labour



Focusing on the solid blue line, the interpretation is that when money wages rise above W_2, further rises in the money wage reduce employment, which is the 'Classical case'. This would arise because aggregate supply (Z) was shifting upwards faster than aggregate demand (D).

How might we explain that? At higher money wages (which might be associated with a higher price level), fiscal and monetary policy might be tightened to head off an inflationary spiral. The resulting negative impact on aggregate demand is likely to reduce the point of effective demand below the previous employment level, N_0 corresponding to wage, W_2.

Further, in an open economy, very high wages might reduce international competitiveness and impinge on export demand, which will also have a negative impact on aggregate demand and shift the point of effective demand to the left of its current position in Figure 14.4.

It is highly probable that the negatively sloped upper segment would be beyond the range defined by normal wage movements and levels. Thus, it is a logical possibility that would be rarely encountered.

Note that the 'Classical' shape is not generated by the dynamics that the Classical system specified (adjustments up a marginal productivity schedule). However, the shape gives the same result: higher money wages reduce employment.

When money wages are below W_0, it is possible that the **Pigou effect** would come into play. This phenomenon, which was named after the British economist Arthur Pigou, is also referred to more generally as the **real balance effect** or the **wealth effect**.

When Keynes attacked the Classical employment theory, he noted that cutting money wages would not be likely to lead to a fall in real wages because competition would also drive prices down, given that firms now enjoyed lower unit costs and assuming productivity did not fall due to low morale brought about by the money wage decline.

In the 1930s, the Classical economists, who had recommended money wage cuts as the way to engineer the real wage cuts considered necessary to restore labour market equilibrium, were forced to acknowledge, albeit reluctantly, that if money wages were cut and prices followed the cost reductions, then the real wage might not fall at all. It was possible that the real wage could even rise if the fall in money wages was less than the fall in prices.

However, Arthur Pigou responded in a famous 1943 article with a proposed solution to the problem of the economy being stuck in an unemployment impasse. He argued that real consumption spending was also a positive function of the stock of the real wealth that individuals possessed. This wealth was held (in nominal terms) in the form of money balances and other financial assets such as government bonds.

Thus, even if a fall in money wages leads to an equivalent percentage fall in the price level, leaving the real wage unchanged, the lower prices would increase the **real** wealth of all those who were holding nominal wealth balances. So, all wealth holders would feel richer, and it was argued, would thus increase real consumption at each level of income.

The increase in real balances at lower prices thus gave proponents of the Classical employment theory another conduit through which money wage falls could stimulate employment, if real wages did not move. In other words, the inverse relationship between money wages and employment was restored by this real balance effect.

It was pointed out, however, that borrowers would feel poorer when prices fell because the real value of their debt burdens would rise, and using the same logic, this would lead to a reduction in real consumption at each level of income. To some extent this would offset any stimulus to spending that the debt holders might contribute.

If most of the debt was in the form of government bonds, then the net effect of falling wages would probably be larger than if wealth were held in the form of private debts.

Thus, when money wages are very low, Weintraub (1956: 844) wrote that: "those owning 'pennies' become 'millionaires' – a calamitous prospect! – full employment may well be assured."

In the real world, if prices fell so low that a real balance effect of any significant size was generated, then it is likely that the entire banking system would collapse. The reason is that while the nominal liabilities held by the banks (many of which would be loans to households and firms) would not be altered, their real values would rise by so much as to bankrupt most of their borrowers. The mass defaults would in turn cripple the financial system.

The empirical evidence is that in normal price movement ranges, the measured real balance or wealth effect is very small and clearly insufficient to remedy a major shortfall in aggregate demand. So, while the Pigou effect presented a logical possibility, it did not provide the Classical employment theory with the support it required to negate the damaging critique made by Keynes.

Money wage rates between W_0 and W_2 – the 'normal range' in Figure 14.4 – are likely to lead to no change in the point of effective demand and thus the macroeconomic demand curve for labour will be vertical. **For employment to change there must be a change in the level of effective demand**.

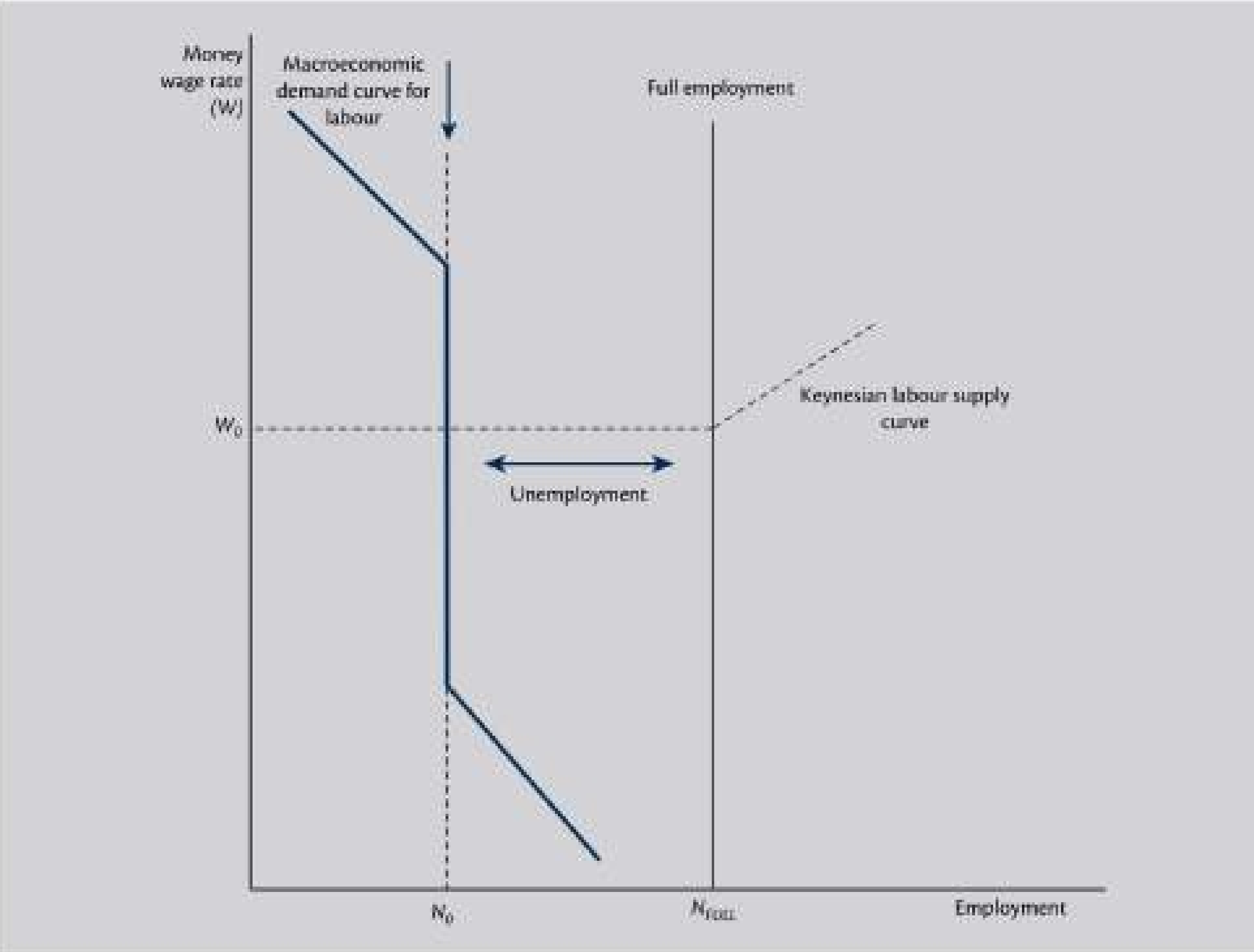
The vertical segment could also be positively sloped if there was evidence of an underconsumptionist response in the normal range of money wage movements. It is possible, for example, in poorer nations that the demand boost from a money wage rate rise will outstrip the supply response arising from the extra production costs. As a consequence, the slope of the macroeconomic demand curve for labour in this relevant range will be positively sloped as depicted in Figure 14.3.

14.3 The Determination of Employment and the Existence of Involuntary Unemployment

We are now able to complete our analysis of the labour market in the macroeconomy by reference to both the demand and supply sides of the labour market.

Figure 14.5 shows the Keynesian labour supply function, which is a function of the money wage. The function tells us that at the prevailing money wage rate (W_0) workers will be willing to supply labour up to the full employment level (N_{FULL}) but after that point they will demand higher wages (to work extended hours and so on). A change in price expectations would lead to shifts in this function.

Figure 14.5 Employment and unemployment



The vertical black line denoted **full employment**, coincides with the employment level at which everyone who desires a job can find one at the prevailing money wage rate (W). The thicker black line is the Macroeconomic demand curve for labour and has backward and forward bending sections at extreme wage levels and a vertical section at normal wage levels.

Involuntary unemployment occurs when the current macroeconomic demand for labour is less than the full employment level. Mass unemployment in the economy is thus determined by the state of effective demand rather than being caused by the ascriptive characteristics of the unemployed themselves. The unemployed become powerless to improve their prospects because the shortage of jobs is caused by a systemic failure of aggregate demand (relative to aggregate supply). In Figure 14.5, the level of involuntary unemployment at this level of effective demand is measured by the distance $N_{full} - N_0$.

The lesson that Keynes taught us was that at an unchanged wage rate a demand stimulus in the goods and services market (which shifts the macroeconomic demand curve for labour outwards towards full employment) would not only stimulate employment. It would at the same time reduce unemployment without any change in the money wage (or price level) being required. This mechanism had been denied by the Classical theory of employment.

Mass unemployment is always driven by insufficient effective demand and the policy prescription is straightforward. For a given level of non-government spending (consumption, investment and net exports), mass unemployment arises because the government spending is too low or taxes are too high. In the event of non-government spending being too low, the cure for mass unemployment is to expand government spending and/or cut taxes to raise aggregate demand.

BOX 14.1 ***THE TALE OF 100 DOGS AND 95 BONES***

Imagine a small community comprising 100 dogs. Each morning they set off into the field to dig for bones. If there are enough bones for all buried in the field, then all the dogs could succeed in their search.

Now imagine that one day the 100 dogs set off for the field as usual but this time they find there are only 95 bones buried.

Some dogs who were always very skilled at finding bones might dig up two bones and others will dig up the usual one bone. But, as a matter of accounting, at least five dogs will return home bone-less.

Now imagine that the government decides that this is unsustainable and decides that it is the skills and motivation of the bone-less dogs that is the problem. They are not skilled or motivated enough. Thus, if the problem were to be constructed to be an individual one, then an individualised solution would be appropriate.

So, a range of dog psychologists and dog trainers might be called in to work on the attitudes and skills of the bone-less dogs. The dogs undergo assessment and are assigned case managers. They are told that unless they train they will miss out on their nightly bowl of food that the government provides to them while bone-less. They feel despondent.

After running and digging skills are imparted to the bone-less dogs, things start to change. While the training helps some dogs improve their luck at finding bones, others turn up boneless. All the training does is to shuffle the queue, always leaving at least 5 dogs without bones.

No amount of training and motivational speeches can resolve the problem; the only solution is to provide more bones.

The point is that when there are insufficient jobs available in the economy, the unemployed are powerless to redress that shortage no matter how hard they search.

Supply side programmes, concentrating on the motivation or skills of the unemployed, will only shuffle the jobless queue in a situation of jobs shortage.

Source: Based on ideas within Centre of Full Employment and Equity (CofFEE) (c.2001)

This was an important lesson that governments learned in the 1930s as a result of the work of Keynes and others. The Classical theory of employment distracted policymakers from seeing that the fundamental solution to unemployment was to increase aggregate demand relative to aggregate supply. As a consequence, in the early years of the Great Depression, millions of workers lost their jobs as governments tried to implement the wage-cutting solutions proposed by the dominant Classical viewpoint.

It was only when governments expanded their deficits that the Great Depression came to an end.

14.4 A Classical Resurgence Thwarted

The fallacy inherent in the Classical faith in wage and price adjustments was first noted by Karl Marx (1863) in his *Theories of Surplus Value* where he discusses the problem of realisation of sales when there is unemployment. He was the first to understand the notion of effective demand. He made the distinction between a **notional demand** for a good (a desire) and an **effective demand** (one that is backed with ability to pay).

It is obvious that the unemployed want to consume more, but because they have no or little income, they cannot translate their notional desires into effective spending. Accordingly, the market, which relies on consumers entering shops with money to purchase goods and services, fails to receive any demand signal from the unemployed and so firms cannot respond with higher production.

This distinction between notional and effective demand was at the heart of the 'Keynes and Classics' debate during the Great Depression. It is central to Keynes' attack on Say's Law, which claimed that 'supply creates its own demand'.

As we have learned, Say's Law denies there can ever be overproduction and unemployment. If consumers decide to save more, then the firms react to this and produce more investment goods to absorb the saving. There is total fluidity of resources between sectors, and workers are simply shifted from making, say, iPads to making investment goods.

Keynes showed that when people save, they do not spend. Further, they give no signal to firms about when they will spend in the future and what they will buy then. So, there is a market failure. Firms react to the rising inventories and cut back output, unable to deal with the uncertainty.

There was a theoretical push to reassert Say's Law using the real balance effect as the conduit by which aggregate demand would always adjust to aggregate supply in the 1950s. But major theoretical work by Keynesians such as Robert Clower (1965) and Axel Leijonhufvud (1968) provided new insights into how we can see the contribution of Keynes and his demolition of Classical theory. These two authors demonstrated in different ways how neoclassical models of optimising behaviour were flawed when applied to macroeconomic issues such as mass unemployment.

Clower (1965) showed that an excess supply in the labour market (unemployment) was not usually accompanied by an excess demand elsewhere in the economy, especially in the product market. Excess demands are expressed in money terms. How could an unemployed worker (who had notional or latent product demands) signal to an employer (a seller in the product market) their demand intentions? Not via saving in financial terms, such as holding money and other liquid financial assets.

Leijonhufvud (1968) also noted that involuntary unemployment arises because there is no way that the unemployed workers can signal that they would buy more goods and services if they were to be employed. A particular firm cannot assume their profit will rise if they hire another worker even though revenue in general will clearly rise (because there will be higher incomes and higher demand). The market signalling process thus breaks down and the economy stagnates. Only if the additional workers hired were guaranteed to buy the firm's extra output could the firm go forward to hire them, because only in that case could the firm be sure that the costs of hiring the workers would be covered by additional sales revenue.[7]

Conclusion

One of the key elements of Keynes' attacks and his ultimate discrediting of the Classical employment theory was his identification of what we call the **fallacy of composition**. We discussed this concept in Chapter 2.

Prior to the 1930s, there was no separate study called macroeconomics. The dominant theory of the day, characterised by the Treasury View, considered macroeconomics to be an exercise in the aggregation of individual relationships. The economy was thus seen as being just like a household or single firm, only bigger. Accordingly, changes in behaviour or circumstances that might benefit the individual or the firm were automatically claimed to be of benefit to the overall economy.

The insistence in the Treasury View that wage cuts would cure the mass unemployment that arose during the 1930s' Great Depression symbolised the fact that their reasoning was based on compositional fallacies.

Keynes led the attack on the mainstream by exposing several fallacies of composition. While these types of logical errors pervade mainstream macroeconomic thinking, there are two famous fallacies of composition in macroeconomics: (a) the paradox of thrift; and (b) the wage-cutting solution to unemployment.

Our discussion of the macroeconomic demand for labour curve in this chapter also highlights how Classical employment theory was bedevilled by this problem.

We have considered the Keynes versus Classics debate in some detail in Chapters 12–14 because the same ideas are in dispute in the current era.

The mainstream response to the persistent unemployment that has beleaguered most economies for the last three or more decades is to invoke supply side measures: wage cutting, stricter activity tests for welfare entitlements, relentless training programmes. But this policy approach, which reflects an emphasis on the labour market, and particularly the wage rate, falls foul of the fallacy of composition problem.

Policymakers consistently mistake a systemic failure for an individual failure. The main reason that the supply side approach is flawed is that it fails to recognise that unemployment arises when there are not enough jobs created to match the desire to work of the willing labour supply. That requires a system-wide policy response to increase effective demand rather than an individual solution focusing on the characteristics of the unemployed.

References

Clower, R. (1965) "The Keynesian Counter-Revolution: A Theoretical Appraisal," in F.H. Hahn and F.P.R. Brechling (eds), *The Theory of Interest Rates*, London: Macmillan.

Centre of Full Employment and Equity (CofFEE) (c.2001) *The Tale of 100 Dogs and 95 Bones*. Available at: http://e1.newcastle.edu.au/coffee/education/education_view.cfm?ID=1, accessed 10 July 2018.

Keynes, J.M. (1936) *The General Theory of Employment, Interest, and Money*, London: Macmillan, 1957 Reprint.

Leijonhufvud, A. (1968) *On Keynesian Economics and the Economics of Keynes: A Study in Monetary Theory*, New York: Oxford University Press.

Marx, K. (1863) *Theories of Surplus Value*. Available at: http://www.marxists.org/archive/marx/works/1863/theories-surplus-value/index.htm, accessed 17 May 2017.

Pigou, A. (1943) "The Classical Stationary State", *The Economic Journal*, December, LIII, 343–51.

Weintraub, S. (1956) "A Macroeconomic Approach to the Theory of Wages", *The American Economic Review*, 45(5), December, 835–56.

Endnote

1. Even in that case, firms might decline to hire the workers. Simply covering the wages paid through sales to the new workers would not generate profits for the firms.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

15 THE AGGREGATE EXPENDITURE MODEL

Chapter Outline

Learning Objectives

- Acknowledge the distinction between identities and behavioural relationships.
- Understand the components of aggregate demand and their determinants.
- Analyse how the components of aggregate demand interact to determine equilibrium national income.
- Understand why the multiplier effect typically exceeds unity.

15.1 Introduction

In Chapter 4, we outlined the measurement of GDP via the expenditure approach in the National Accounts (NIPA), underlining the fundamental premise in macroeconomics that it is total spending that drives output (GDP) and employment in the economy. However, National Accounts principles are founded on definitions which give rise to an identity equating total spending to GDP and to NI (national income).

In this chapter, we develop a model of output (GDP) determination, which will provide us with an understanding as to how the different components of expenditure interact and determine total output. This means we must outline the key variables which influence the components of total spending so that we can develop behavioural relationships. These components of expenditure are: consumption, investment, government expenditure and net exports.

In Chapter 13, we introduced the concept of effective demand within the original aggregate supply-aggregate demand framework (D-Z) developed by Keynes in *The General Theory of Employment, Interest and Money* (1936). Recall that Keynes defined the aggregate supply price of the output derived for a given amount of employment as the "expectation of proceeds which will just make it worth the while of the entrepreneur to give that employment" (Keynes, 1936: 24).

We learned that this concept relates a volume of revenue received from the sale of goods and services to each possible level of employment. At each point on the aggregate supply price curve, the revenue received would be sufficient to cover all production costs and desired profits at the relevant employment level.

The important point here is that the relationship is expressed in terms of employment and the revenue expected to be received at each output level. We should be clear that Keynes was considering aggregate supply in terms of the expectation of proceeds, which is the money income (revenue) that the firms expect to get from selling their output. Firms are thus considered to formulate plans to gain a volume of money or nominal profits.

While GDP as a measure of output is fraught with problems, as we discussed in Chapter 4, the difficulty for our current purposes is that GDP can rise simply because prices rise. What Keynes wanted was a consistent measure of economic activity. It is for that reason that Keynes used employment instead of output as his measure of aggregate activity.

In this chapter, we thus have two tasks. First, we must express the model developed in Chapter 13 in terms of output rather than employment. Second, we consider total expenditure and relate that to a simple model of aggregate supply. In Chapter 16 we will complete the picture by developing a more detailed aggregate supply framework. We can do that by using the National Accounting concept of GDP at constant rather than current prices as our measure of economic activity.

In this way, we do not diminish the importance of expectations in driving production and supply decisions by firms, but we abstract from price changes and assume that firms react to changes in aggregate spending by adjusting only the quantity of output rather than both the price and the quantity. In other words, we adopt for the time being, the assumption that firms respond dollar for dollar to increased demand by increasing output and income. We can justify this assumption of **quantity adjustment** in two ways.

First, firms use mark-up pricing principles, whereby they add a profit mark-up to their unit costs and face roughly constant unit costs over the output range within which they normally produce. Typically, they maintain some excess capacity and can thus increase output relatively easily without further investment in productive capacity, which would take time. If they face insufficient capacity relative to demand, firms are likely to raise prices to ration demand, which inevitably leads to the loss of customers to competitors.

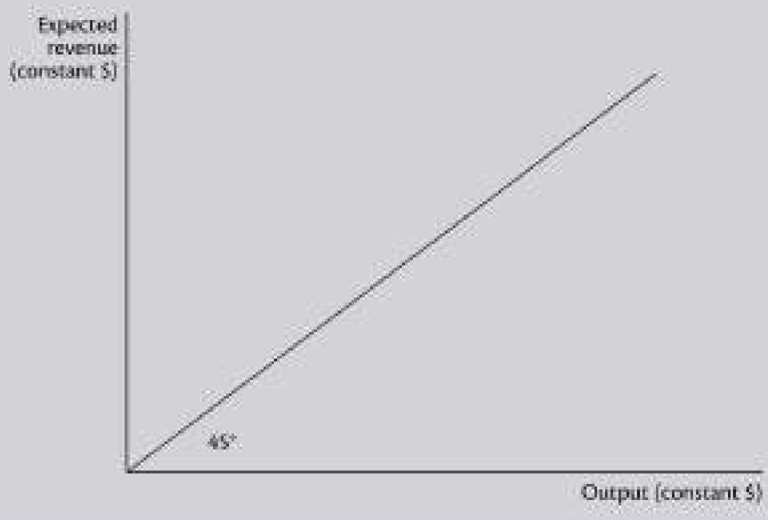
Second, firms face various costs when adjusting prices and thus only periodically make such adjustments. It has been said that firms use 'catalogue pricing', whereby they make their prices known to their prospective customers through advertising and other means and then are prepared to sell goods and services at those prices irrespective of demand (up to their full productive capacity). At the end of the current catalogue period, they will then make any necessary adjustments to prices based on expected future demand and any recent and expected movements in unit costs (see Chapter 16).

15.2 A Simple Aggregate Supply Depiction

Figure 15.1 depicts the simplified **constant price aggregate supply** relationship that we will work with in this chapter to concentrate on the way the economy adjusts in quantity terms only to changes in aggregate demand. The relationship is drawn as a 45° line emanating from the origin with total expected revenue on the vertical axis and output on the horizontal axis.

The 45° line shows all points where the expected revenue is equal to output (aggregate supply). We also measured planned expenditure in constant prices (inflation-adjusted aggregate demand).

Figure 15.1 Aggregate supply



In Chapter 4 we explained the different perspectives that we can take in measuring aggregate economic activity. The expenditure, income and output approaches provide different views of the national accounting framework but they yield the same aggregate outcome. The total value of goods and services produced in any period is equal to the total spending and the total income generated (wages, profits, rent and profit) in that same period.

As we will see in Chapter 16, when we consider aggregate supply in more detail, firms supply a particular level of output (and incur costs of production) as long as they can generate enough revenue to cover the costs and realise their desired profits.

Note that the vertical axis shows the revenue that is *expected* to be generated by employing workers and producing output. In other words, it is forward looking. However, by producing output, incomes will be generated, most of which will be spent. The 45° line shows us the points where the revenues generated are equal to the expenditures, which of course must be true because spending must generate revenue.

The expected revenue provides a measure of total revenue or expenditure, which also equals income. From the perspective of firms, it tells them the expected proceeds that can be generated by selling the different levels of output.

The other point to note is that we have not represented full capacity output on the graph. When the economy is operating at full capacity, firms are unable to continue expanding output in response to additional spending. When we formally introduce the expenditure side of the economy in the next section, we will also define full employment output beyond which firms cease to be quantity adjusters.

15.3 Aggregate Demand

In Chapter 4 we learned that firms generate additional productive capacity through new investment in order to produce additional goods and services to satisfy demand. Here for simplicity, we are assuming no depreciation of the existing stock of productive capacity. Once the capital stock is in place, firms will respond to increases in spending for the goods and services they supply by increasing output up to the productive limits of their capital and the available labour and other inputs. Beyond full capacity, they can only increase prices when increased spending occurs. We assume that potential output is fixed for the period we are analysing.

By adopting the assumption that prices do not adjust to changes in demand, we thus say that higher aggregate demand will lead to increased production, which in turn increases national income.

Note that as a result of our fixed price assumption, all extra expenditure increases both GDP and constant price GDP (GDP adjusted for inflation).

The basic macroeconomic rule is that, subject to the existing productive capacity, total spending drives output and national income, which in turn, drives employment.

In Chapter 4 we learned that total expenditure in any period is expressed as the sum of the following sources of spending:

- Consumption by households or persons (C).
- Investment spending by firms (I).
- Government spending (G).
- Export spending by foreigners (X) minus import spending by domestic residents (M), which we denote as net exports, $NX = (X - M)$.

The sum of these expenditures equals GDP as a matter of accounting. Total expenditure sums to total output and also total income.

From the National Accounting framework, we know that total expenditure (E) in the domestic economy in any particular period can be expressed as:

$$(15.1) \quad E = C + I + G + (X - M)$$

You will see that while exports add to total spending in the domestic economy, imports lead to a drain in spending because they represent the spending of local residents, firms and governments on goods and services produced by other nations.

The equilibrium level of national income (Y) is determined by aggregate demand:

$$(15.2) \quad Y = E$$

You should note that the level of GDP that is produced by the current period's expenditure does not necessarily have to equal the full employment output level. Keynes among others demonstrated that full employment was not guaranteed by the market system.

We considered those issues in detail in Chapters 5 and 12 when we studied the labour market.

In the remainder of this chapter we will develop a more detailed understanding of the behaviour of each of these components of total spending and explain how they interact to determine total output (GDP) and national income. We will also derive an expression for equilibrium national income.

15.4 Private Consumption Expenditure

To gain an understanding of the determinants of aggregate demand, we have to focus on private sector decision making which occurs within a broad set of constraints resulting from the interaction between the government and non-government sectors.

Private consumption spending is the largest component of total spending on GDP in most economies. Consumption is the sum of household spending on non-durable goods (for example, food), durable goods that provide benefits beyond a single year (for example, cars and white goods such as refrigerators) and services (for example, restaurants, theatres and the like).

Table 15.1 shows the ratio of private consumption expenditure to total GDP for most of the OECD nations. While there are notable exceptions, the outcomes in most nations are close to the OECD average of 60.8 per cent. The ratios across countries are also relatively stable over time.

What determines total private consumption expenditure?

The most elementary theory of private consumption (C) is that it is a stable function of disposable national income (Y_d). Disposable income is the flow of income that remains after taxes (T) have been paid.

Table 15.1 Consumption ratios, OECD nations, 2010 and 2016, per cent

	2010	2016
Australia	54.8	56.8
Austria	51.5	50.5
Belgium	50.7	50.0
Canada	55.6	56.7
Czech Republic	48.3	46.4
Denmark	46.1	46.0
Estonia	50.8	51.1
Finland	50.5	52.6
France	53.4	52.3
Germany	54.5	51.5
Greece	67.3	67.1
Hungary	51.0	48.1
Iceland	49.6	47.3
Ireland	46.5	32.2
Italy	60.5	60.3
Japan	56.5	54.2
Korea	48.6	46.3
Latvia	62.7	60.3
Lithuania	63.8	64.1
Luxembourg	30.8	28.8
Mexico	63.8	64.3
Netherlands	43.9	43.4
New Zealand	56.8	56.2
Norway	40.0	43.1
Poland	60.6	57.8
Portugal	63.9	63.4
Slovak Republic	56.8	53.6
Slovenia	55.1	52.6
Spain	56.2	56.6
Sweden	45.0	42.9
Switzerland	52.2	51.9
UK	62.3	62.9
US	66.3	66.9
OECD average	59.7	59.3

Source: Authors' own. Data from OECD.Stat, Annual National Accounts.

We define disposable income as:

$$(15.3) \quad Y_d = Y - T$$

A simple model of the government's tax policy is that it levies a proportional tax rate (t) on total national income, which means that the total tax revenue at income, Y, is given as:

$$(15.4) \quad T = tY$$

Assume that the proportional tax rate (t) is 0.20. This means that for every dollar of national income generated in the economy the government takes 20 cents out in the form of taxation. The 80 cents that remains, is disposable income. If national income was equal to \$1,000 then the total tax receipts accruing to government would be equal to \$200.

We consider taxation to be a **leakage** from the expenditure system because it is income that is not available for private spending to be recycled back into the economy.

Disposable income can be written as:

$$(15.5) \quad Y_d = Y - T = Y - tY = (1 - t)Y$$

In our example, this would be written as $Y_d = (1 - 0.2)Y = 0.8Y$.

In macroeconomics, the **aggregate consumption function** is defined as a relationship between total consumption (C) and total disposable income (Y).

$$(15.6) \quad C = C_0 + cY_d$$

C_0 is a constant and is a base level of consumption which is independent of disposable income and is therefore called autonomous consumption. The coefficient c is called the **marginal propensity to consume** (*MPC*) and measures the fraction of every additional dollar of disposable income that is consumed.

The *MPC* is generally presumed to have a value between zero and one. If, for example, $c = 0.75$, then for every extra dollar of disposable income that the economy generates, consumption would rise by 75 cents.

It is important to understand that the *MPC* in this model is an aggregate, which is an average of all the individual household consumption propensities. Lower-income households tend to have *MPC* values close to one, whereas the higher-income households have much lower than average consumption propensities.

This arises because lower-income families find it harder to purchase enough goods and services to maintain basic survival. Higher-income earners not only consume more in absolute terms, but also have more unspent income after they have purchased all the basic essentials and luxury items.

As we will learn later in this chapter, the distribution of income is an important consideration when seeking to understand changes in aggregate demand. For example, a change in tax policy that increased disposable income for low-income consumers would have a greater positive impact on final consumption than a tax cut aimed at giving high-income earners the same absolute increase in disposable income.

By substituting the expression for disposable income in Equation (15.5) into the consumption Equation (15.6), we can show the direct influence of the tax rate on private consumption:

$$(15.7) \quad C = C_0 + cY_d = C_0 + c(1 - t)Y$$

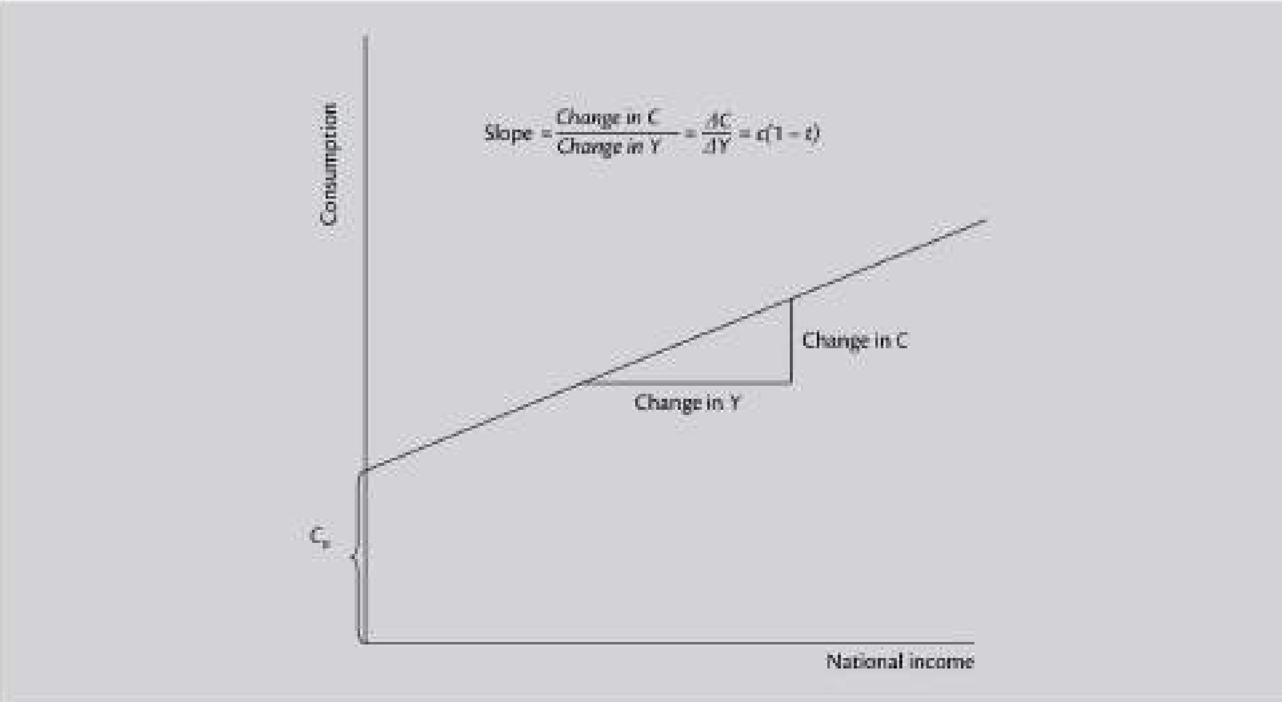
Figure 15.2 shows the consumption function in graphical terms. Note here and in the graphs to follow that the vertical axis is specified in terms of actual expenditure (rather than expected expenditure (or revenue from the perspective of firms) as in Figure 15.1). For simplicity, we will assume that expected and actual revenue are equal, although in the real world that is not likely to be the case. Later we will consider what happens when expectations are not met.

The consumption function cuts the vertical axis at $C_0 > 0$. The consumption function is then upward sloping because we have postulated that consumption rises with national income.

In Chapter 7 *Methods, Tools and Techniques*, we learned how to derive a slope graphically. The slope of a line is the ratio RISE over RUN. Rise in this case is the change in consumption spending (ΔC) and run is the change in national income (ΔY) as shown in Figure 15.2.

In fact, $\Delta C = c(1 - t)\Delta Y$ and RISE over RUN $= \Delta C/\Delta Y = c(1 - t)\Delta Y/\Delta Y = c(1 - t)$. The slope of the consumption function is thus given by the coefficient $c(1 - t)$, which is lower than the *MPC* because a one dollar rise in national income translates into less than a one dollar rise in disposable income, given that the marginal tax rate (t) is positive.

Figure 15.2 The consumption function



You should be able to work out what would happen if the MPC (c) increased. The result would be an increase in the slope of the consumption function such that at every level of disposable income, total consumption would be higher.

What determines aggregate saving (S)?

$$S = Y_d - C = Y - T - C \quad (15.8)$$

Saving at the macroeconomic level is a **residual** that remains after households have made their spending decisions.

Given the MPC out of disposable income takes the value (c), we can define a related concept; the **marginal propensity to save** (MPS, or s) which is one minus the marginal propensity to consume: $s = 1 - c$. Note that since disposable income can only be consumed or saved, $1 = s + c$ or MPS + MPC = 1.

When national income rises, the government takes out some taxes, leaving an increase in disposable income which is then the source of increased consumption (via the marginal propensity to consume), with the remainder of the increase in disposable income being saved.

If the MPC is 0.75 and the tax rate is 0.2, then if national income increases by $1,000, total tax revenue rises by $200, so disposable income rises by $800. Consumption then rises by 0.75 × $800 = $600 and $200 is saved.

15.5 Private Investment

When macroeconomists use the term 'investment' they are referring to a very specific type of spending that does not accord with the common usage of the term. For example, a lay person might think of investment as a person putting some money in a fixed term deposit at a bank or purchasing some shares in a company.

The National Accounting meaning of investment is any spending that adds to the productive capacity of the economy, that is, adds to the **capital stock**. Capital in this context is productive plant and equipment or other capacity, which defines the **potential output** of an economy. Thus, when a firm builds a new factory or purchases a new piece of machinery, it is considered to be investing. (If you are having trouble distinguishing between the two, ask yourself this question: does the investment require labour to produce it? Clearly it takes labour to produce plant and equipment, while it does not require labour to produce a checkable deposit at the bank; aside from the insignificant amount of labour to make an entry on the bank's balance sheet.)

Generally, firms invest, but households consume and save. The one major exception is residential real estate investment, which is included as investment. But here again, the economists' definition differs somewhat from the layperson's because only *newly constructed* housing counts. Purchases of existing ('used') housing do not count as investment for the purposes of GDP accounting because it is not newly produced. (Again, if you buy existing housing, there is no labour required at that time to produce it; housing is counted as investment only for the period in which it was built.)

Changes to the stock of inventories, which are unsold goods, are also considered to be a component of business investment in each period because they add to the potential of the economy to meet current aggregate demand for goods and services. As we will see, the dynamics of inventories provide important information about the state of the business cycle.

Table 15.2 presents a snapshot of the Australian National Accounts, showing the quarterly and annual growth rates of the components of GDP up to the final quarter of 2017, which are based on the seasonally adjusted expenditure **chain volume measures**. The organisation of the expenditure components is based on the standard National Accounts framework that we discussed in Chapter 4, and which is broadly shared across the world. Under the heading *Gross fixed capital formation*, are listed a number of individual line entries which comprise the separate categories that the statistician uses to estimate total investment spending. The *Change in inventories* is classified as a separate category.

Economists distinguish between **gross investment** and **net investment**. Gross investment is the total spending by firms on new plant and equipment and on inventories. However, in each period the existing capital stock depreciates. For example, machines wear out and/or become obsolete; buildings require maintenance; and car fleets require updating. Some of the gross investment in each period merely covers the depreciation of the existing capital stock.

Table 15.2 Expenditure chain volume measures in national accounts (seasonally adjusted), Australia, 2017

	Percentage change September 2017 to December 2017	Percentage change December 2016 to December 2017	Percentage points contributions to growth in GDP September 2017 to December 2017
Final consumption expenditure			
General government	1.7	4.9	0.3
Households	1.0	2.9	0.6
Gross fixed capital formation			
Private			
Dwellings	−1.3	−5.8	−0.1
Ownership transfer costs	−2.1	1.5	–
Non-dwelling construction	−8.0	8.0	−0.5
Machinery and equipment	3.3	8.7	0.1
Cultivated biological resources	−1.4	3.3	–
Intellectual property products	0.6	4.0	–
Public	2.9	1.5	0.2
Changes in inventories	NA	NA	–
Gross national expenditure	0.6	3.0	0.6
Exports of goods and services	−1.8	0.8	−0.4
Imports of goods and services	0.5	6.6	−0.1
Statistical discrepancy (E)	NA	NA	0.2
Gross domestic product	0.4	2.4	0.4

Source: Data from Australian Bureau of Statistics (2018).

Net investment is the component of gross investment that adds **new** productive capacity, that is, increases the overall capital stock. Net investment is thus gross investment less total depreciation. In this chapter, we assume that there is no depreciation of existing capital stock and no planned changes to the stock of inventories. In other words, for simplicity we assume that gross and net investment are equal.

We assume for now that investment is a purely autonomous or exogenous source of spending in the economy, an **injection** of expenditure.

15.6 Government Spending

We have already introduced one element of the government's interaction with the non-government sector, namely the proportional tax rate (t), which is our simplified expression for what is a complex tax structure in real world economies.

The National Accounts framework shows that government spending takes a variety of forms. First, all levels of government purchase a range of goods and services from the non-government sector as a means of fulfilling their social and economic goals. Some of the purchases are for consumption goods and services (yielding services over a fixed period, usually 12 months, such as outsourcing welfare services), while other spending is categorised as public investment or public capital formation (which yield services over many periods, such as a federally funded public transport system). The latter category of spending generates the valuable public infrastructure that enhances the welfare and profitability of the non-government sector.

Second, governments directly employ workers, such as public servants who provide a range of services to the public.

Third, the government provides a range of **transfer payments** to the non-government sector in the form of pension and welfare entitlements and so on. In Chapter 4, we learned that the National Accounts framework did not include transfers as a component of government spending because they do not constitute a final demand for goods and services by government. The final demand impact of the transfer payments is measured when the recipients (for example, households) spend.

In our models, which are stylisations of the National Accounting measures, the flow T (total taxes) represents **net taxes** which is total tax revenue minus total transfers to the non-government sector.

Net government spending (that is, government spending minus taxes net of transfers) is determined by two broad forces.

1. The discretionary decisions that the government takes in setting its fiscal policy (that is, levels of expenditure and tax rate(s)).
2. The state of the overall economic cycle impacts on net taxes and hence net government spending. For example, when the economy is performing badly, net taxes will fall because of both lower taxes and higher welfare payments even without any explicit change in government policy. The opposite will be the case when the economy is growing strongly and unemployment is falling. We call these effects cyclical because they vary with the state of the economic cycle. They are also termed **automatic stabilisers** because they function automatically and act to stabilise total spending in both directions. We briefly discussed automatic stabilisers in Chapters 8 and 14.

For the purposes of the following discussion, we will assume away these cyclical effects on government spending (G) and assume that its level is exogenous to national income. We learned about the meaning of exogenous variables in Chapter 7 *Methods, Tools and Techniques*.

Government spending is thus considered to be an **injection** of expenditure into the economy in contrast to taxation which is a **leakage**.

15.7 Net Exports

Exports are goods and services produced in the local economy, which are then sold to the rest of the world. While export spending boosts national income, we consider exports to be a cost in the sense that they deprive the domestic population of the use of the real resources that are absorbed in the production of the goods and services sold abroad.

Imports comprise expenditure on goods and services which are produced by the rest of the world, by households, firms and government,. In other words, some of the consumption expenditure, investment expenditure and government expenditure in each period does not increase domestic production and is thus considered to be 'lost' because it has 'leaked out' of the domestic expenditure/income loop. Imports are thus **leakages** of expenditure from the economy.

Even though import expenditure is a leakage from the expenditure system, we consider it to provide benefits to the domestic economy by allowing households, firms and government to enjoy access to goods and services that are not otherwise available or are not available on competitive (quality and/or price) terms.

The difference between exports (X) and imports (M) is called the **net exports** (NX) of a nation. A trade surplus occurs when exports are greater than imports. A trade deficit occurs when the opposite is the case. We will consider these issues in Chapter 24, when we will consider the determinants of net exports in detail by introducing the exchange rate and measures of international competitiveness.

In this chapter, we simplify the model by assuming that exports (X) are given in any year and determined by national income in the rest of the world, which is outside the influence of the domestic economy.

What determines import spending?

We assume that import spending (M) rises with national income. Thus, the higher is the national income, the greater will be the flow of imported consumer goods and services and imports of capital equipment.

For now, we assume that a nation imports a fixed proportion of every dollar of national income. That proportion is called the **marginal propensity to import** (m) and has a meaning similar to the MPC. The marginal propensity to import is the extra import spending that occurs as a result of a dollar increase in national income.

Our simplified import expenditure model is given as:

$$(15.9) \qquad M = mY$$

For example, if $m = 0.2$, then if national income (Y) rises by \$1,000, import spending will increase by \$200. The higher is the marginal propensity to import, the higher is the leakage in import spending at every level of national income.

15.8 Total Aggregate Expenditure

Now that we have considered all the components of aggregate expenditure (or demand) we can write the equation for aggregate demand (E) as:

$$(15.10) \quad E = C + I + G + (X - M)$$

As we saw at the beginning of this chapter, this is an accounting statement, which is derived from the National Accounting framework.

We have developed some simple behavioural theories about the individual components of total expenditure. We assumed that consumption and imports were positive functions of national income and that investment, exports and government spending were determined in each period by factors invariant to national income. We also adopted a simple tax rule.

We can expand the aggregate demand equation to reflect those behavioural assumptions, which in turn, will allow us to derive an expression for equilibrium income.

$$(15.11) \quad E = C_0 + c(1-t)Y + I + G + X - mY$$

You can see that some components of total expenditure in the economy are dependent on national income levels and some (I, G and X) are, by assumption, independent of or autonomous to national income.

We can use the techniques in Chapter 7 *Methods, Tools and Techniques* to modify Equation (15.11):

$$(15.12) \quad E = C_0 + I + G + X + [c(1-t) - m]Y$$

We could usefully simplify this expression by adding together all the components of total expenditure that are autonomous (such that $A = C_0 + I + G + X$) to write the **aggregate demand function** as:

$$(15.13) \quad E = A + [c(1-t) - m]Y$$

The slope of this function is $\Delta E/\Delta Y = [c(1-t) - m]$. The slope of the function will be steeper, the higher is the marginal propensity to consume (c) and the lower are the tax rate (t) and the marginal propensity to import (m).

TRY IT YOURSELF

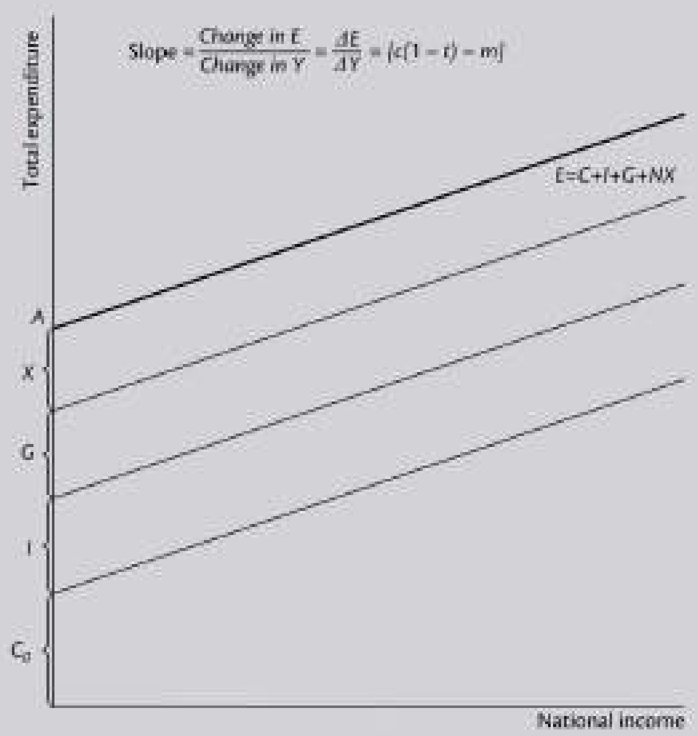
See if you can explain this result to yourself at this point. Think about a positive marginal propensity to consume as leading to an increase in **induced consumption** expenditure when national income rises. The tax rate and marginal propensity to import are leakages from the expenditure system for each dollar rise in national income. We will provide a full analysis of this in the next section when we consider the **expenditure multiplier**.

Figure 15.3 shows the aggregate demand function and its individual autonomous components.

The aggregate demand function is drawn with national output/income on the horizontal axis and aggregate demand or planned expenditure on the vertical axis. It shows the relationship between national income and planned expenditure and two aspects are important:

- Total autonomous spending, A, is the intercept on the vertical axis; and
- The slope of the function depends on the marginal propensity to consume (c), the marginal propensity to import (m) and the tax rate (t).

Figure 15.3 The aggregate demand function



The aggregate demand (planned expenditure) schedule will shift if any of the components that make up autonomous spending change. Also the slope of the schedule will change, if any of *c*, *t* or *m* change.

If planned government spending or private investment were to rise, then the aggregate demand function would shift upwards, with the vertical intercept rising by the increase in planned spending. As a result, planned spending would be higher at each income level. The opposite would be the case if, for example, government spending was to be cut. In each case the slope of the function would be unchanged. Figure 15.4 shows the impact of a rise in autonomous spending from A_0 to A_1.

The slope of the aggregate demand function may also change. A rise in the marginal propensity to consume (c) and/or a fall in the tax rate (t) and/or a fall in the marginal propensity to import (m) will increase the slope of the function.

Figure 15.5 shows the impact of a rise in the marginal propensity to consume (c). At every national income level, total planned spending is now higher.

TRY IT YOURSELF

See if you can sketch the impacts and explain the meaning of a rise in the tax rate (t), a rise in the marginal propensity to import (m) and a fall in the marginal propensity to consume (c).

You should also try to articulate what factors will influence the actual values of the marginal propensity to consume (c) and marginal propensity to import (m) in the real world. For example, poorer nations with basic financial systems might be expected to have a higher marginal propensity to consume than nations with high disposable income.

Figure 15.4 Increase in the intercept of the aggregate demand function with increased autonomous spending

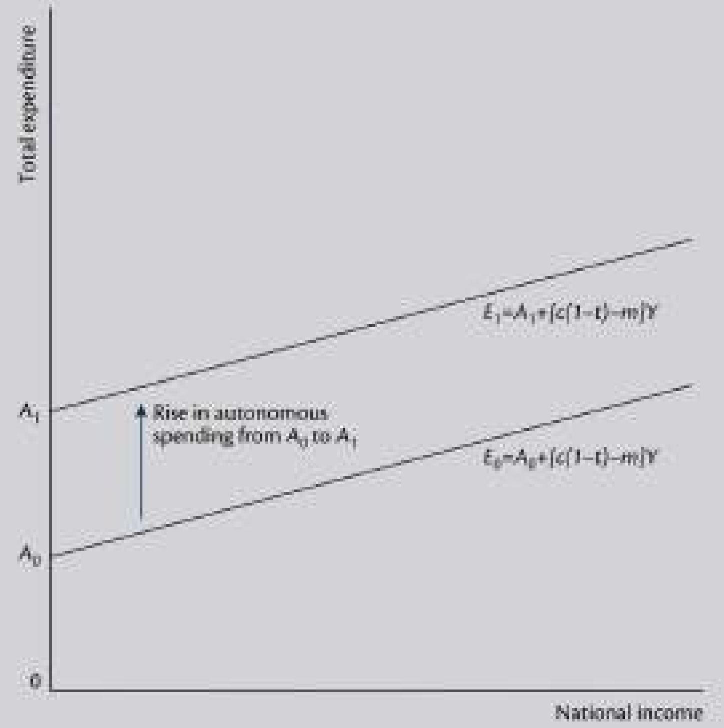
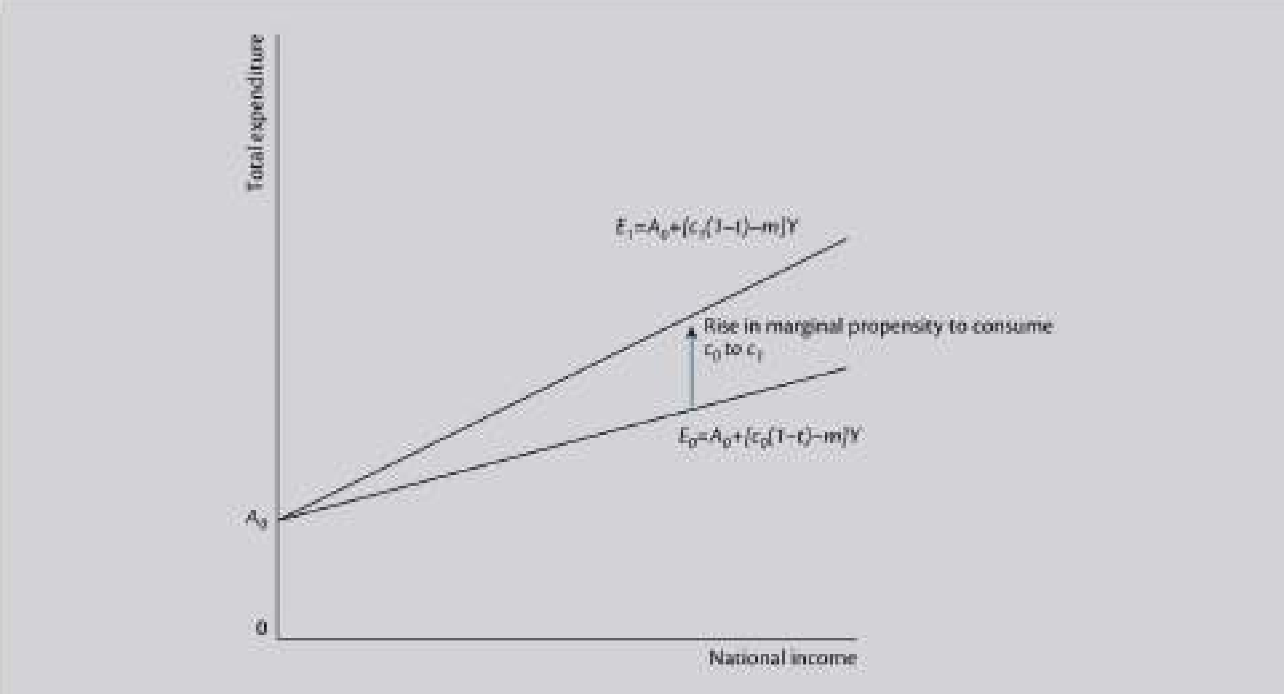


Figure 15.5 Changing slope of the aggregate demand function with increased marginal propensity to consume



15.9 Equilibrium National Income

The term **equilibrium** in macroeconomics is used to refer to a situation where there are no forces present which would alter the current level of spending, output and national income. At that point, firms are selling all the output they produce based on their expectations of planned expenditure. Equilibrium is associated with a position of rest.

You should be very careful not to confuse equilibrium with full employment. A macroeconomic equilibrium can also occur at times when there is very high involuntary unemployment. Full employment is only one possible point of equilibrium.

Equilibrium occurs when planned expenditure is equal to national income and output.

Under our current assumption that firms in the economy are quantity adjusters and prices are fixed in the short term, Figure 15.1 showed us that the 45° line was the aggregate supply curve, since the level of aggregate supply equals national income.

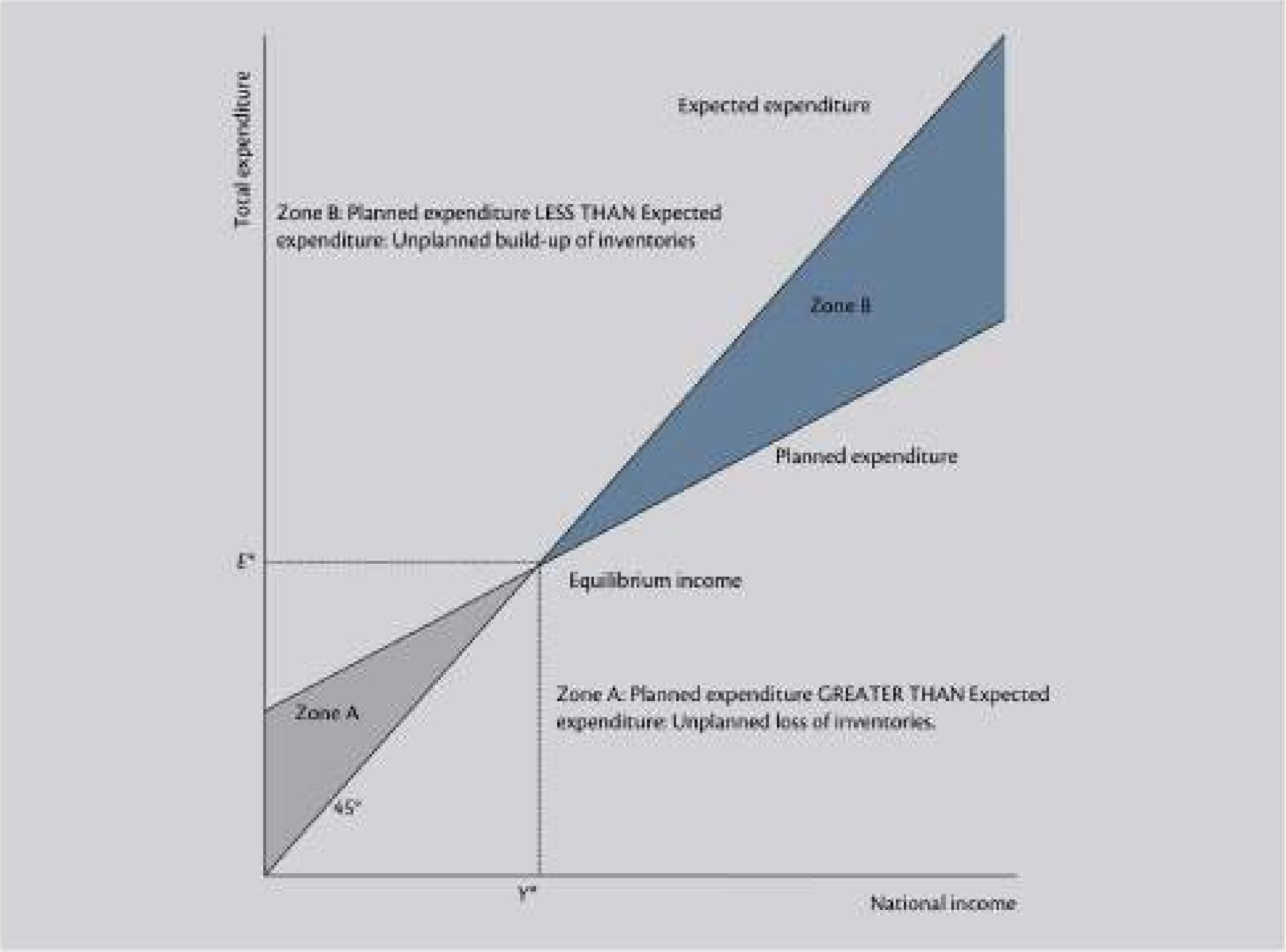
We are implicitly assuming that there are idle resources available for firms to deploy in expanding output. At some point, when the economy is operating at full capacity, firms are unable to continue expanding the quantity of output in response to additional spending. At that point, price rises are inevitable.

Equilibrium thus occurs when the aggregate demand function cuts the 45° line because at this point the aggregate demand expectations formed by the firms, which motivated their decisions to supply (Y^*), are consistent with the total planned expenditure (E^*) of consumers, firms, government and the external economy.

Figure 15.6 shows the equilibrium income (Y^*) and expenditure (E^*) combination that defines the effective demand in the economy at this point in time.

In Figure 15.6, note also the two areas that lie between the aggregate demand function and the 45° line. Zone A is characterised by planned expenditure being greater than expected demand and hence actual output and income. Firms have supplied insufficient output and generated less national income than would be consistent

Figure 15.6 Planned expenditure and equilibrium income



with actual planned expenditure. In this situation, there is an unplanned reduction in the stock of inventories, which provides the signal to firms that they have been mistaken in their expectations. Firms would react to this unplanned run-down in inventories by increasing output and national income.

On the other hand, Zone B is characterised by planned expenditure being less than expected demand and hence actual output and income. Firms have been too optimistic and oversupplied output, generating more national income through their production of output than would be consistent with actual planned expenditure. In this situation, an unplanned increase in the stock of inventories provides the signal to firms that they have been mistaken in their expectations. Firms would react to this increase in inventories by decreasing output and national income.

The inventory cycle is an important part of the cyclical adjustments that quantity-adjusting firms make to bring their expectations and production decisions into line with planned expenditure, which brings the economy back into equilibrium. The decision to increase production means that more employment will be created and the higher national income leads to an increase in planned expenditure. Firms will continue increasing output, income and employment until their expectations are matched by planned expenditure and there are no further unplanned reductions in inventories. Equilibrium is reached at the Y^*-E^* combination, that is, where the aggregate demand function cuts the 45° aggregate supply line.

Similarly, if firms find that they have produced too much, so that there is undesired inventory, they will cut back employment and production. Firms will continue decreasing output, income and employment until there are no further unplanned increases in inventories and the Y^*-E^* combination is achieved. At that point, planned inventories are being held and firms are producing in line with planned expenditure.

BOX 15.1 INVENTORY MOVEMENTS AND PLANNED INVESTMENT

We have learned that unplanned changes in the stock of inventories lead to GDP and national income adjustments because they signal to firms that their past expectations with respect to current aggregate demand, on which they based their current production decisions, were wrong.

If the stock of inventories starts to increase beyond the normal level that firms maintain to meet the flux in spending, it signals that firms were overly optimistic about the level of aggregate demand. Once they form the view that the discrepancy is not a random event, they will cut back on production and national income will fall.

Conversely if inventories start to be depleted below the normal level and firms think this is not an ephemeral episode, then GDP will rise because firms will revise their expectations of aggregate demand upwards. Output, employment and national income will rise as a result.

We have defined national income equilibrium as occurring when planned aggregate demand equals GDP or national income. In Chapter 4, we learned that in the national accounts aggregate spending always equals GDP or national income, irrespective of whether there is national income equilibrium.

However, the accounting concept of total expenditure is slightly different to our macroeconomic concept of planned aggregate demand. The difference is that the flow of spending on inventories in any given period need not accord with the planned expenditure on inventories that firms choose to make to meet the normal fluctuations in their sales.

The way the national accounts deal with this discrepancy is to classify all inventory expenditure in a period as part of gross capital formation or investment.

In macroeconomics, we conceptualise the discrepancy by differentiating between planned (*p*) and unplanned (*u*) aggregates. So total investment, $I = I_p + I_u$, where I_p and I_u are the planned and unplanned investment, including build-ups (or losses) of inventories, respectively. It is the unplanned accumulation (or decrease) that is associated with changes in GDP and national income. Thus, I_u can be positive(negative), so if sales are lower(higher) than expected, the stock of inventories rises(falls) and total investment will be higher(smaller) than planned.

So, in the national accounts for a period, a non-zero level of I_u would be included as part of inventory investment. But from a macroeconomic theory perspective, we would consider a positive or negative value for I_u in any period as providing evidence that firms' expectations have been inaccurate and there is a dynamic process in the economy whereby firms change output, GDP and national income. Equilibrium thus implies that $I_u = 0$.

15.10 The Expenditure Multiplier

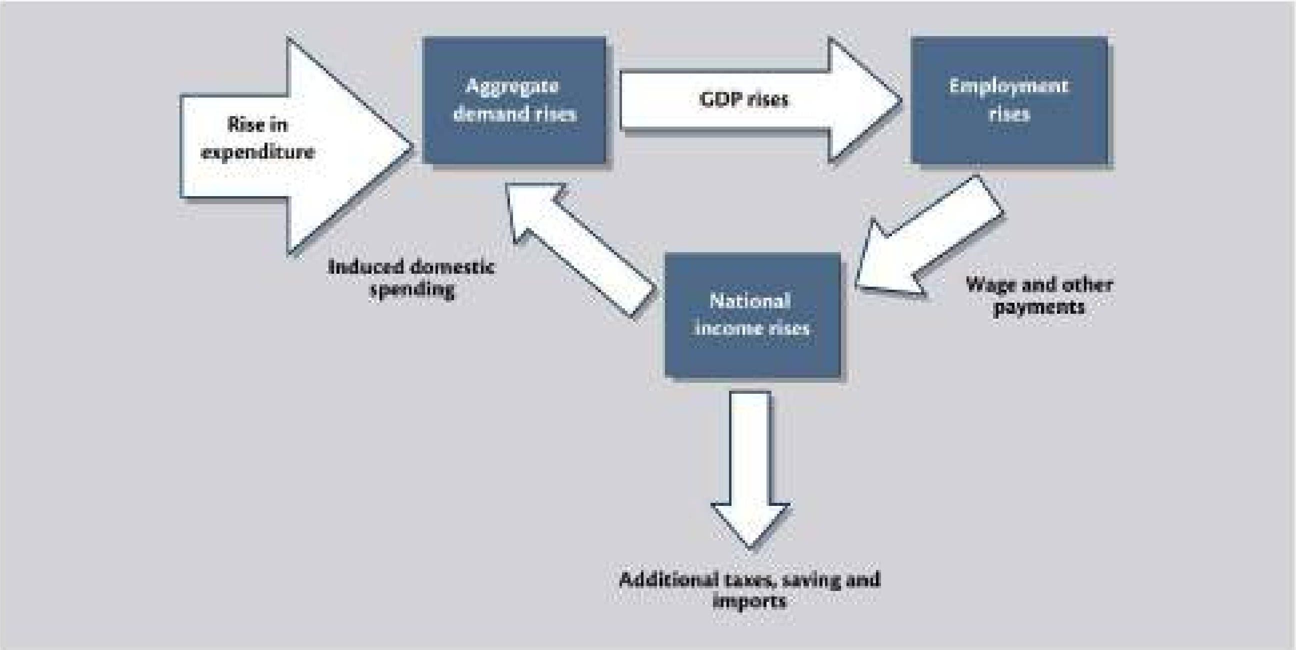
There is an additional feature of the income adjustment mechanism that is important to understand. The aggregate demand function (total planned expenditure) in Equation 15.13 is composed of two components: (a) the autonomous spending component ($A = C_0 + I + G + X$); and (b) the expenditure induced by the level of national income, $[c(1 - t) - m]Y$.

What would be the impact on GDP (and national income) if one of the components of autonomous spending changed? We know that GDP and national income will rise if planned spending rises and will fall if planned spending falls. The question of interest now is by how much will GDP and national income change after a change in planned spending that is driven by a change in autonomous spending (for example, an increase in government spending).

Economists have developed the concept of the **expenditure multiplier** to estimate how much national income (Y) will change for a given change in autonomous spending (A).

Figure 15.7 sketches the expenditure multiplier process. From an initial equilibrium position, an increase in autonomous expenditure provides an instant boost to aggregate demand. Firms respond to the increased

Figure 15.7 The multiplier process



planned expenditure and raise employment to produce the increased output (GDP), as explained in the previous section. National income increases.

This rise in national income induces further consumption spending which leads to a further rise in aggregate demand, employment and GDP. A proportion of the rise in national income leaks out in the form of higher tax payments and imports and increased saving.

The process continues until the induced spending becomes so small that there are no further GDP increases. On the other hand, a fall in autonomous expenditure leads to a series of cuts in employment, GDP and tax payments, imports and saving.

An algebraic treatment

The formal expression for the expenditure multiplier is derived directly from the equilibrium national income and expenditure relationship.

The aggregate demand function was expressed as $E = A + [c(1 - t) - m]Y$. The national income equilibrium condition is given as:

(15.14) $Y = E$

If we substitute the equilibrium condition into the aggregate demand function (15.14) we get:

(15.15a) $Y = E = A + [c(1 - t) - m]Y$

and solving for Y (by collecting Y terms on the left-hand side) gives:

(15.15b) $Y[1 - c(1 - t) + m] = A$

Thus equilibrium income is:

(15.15c) $Y = A/[1 - c(1 - t) + m]$

The expenditure multiplier (α) is defined as the change in national income with a change in autonomous spending ($\Delta Y/\Delta A$) and, using differential calculus, you can see that it is equal to the coefficient of the A term in Equation (15.15c):

$$(15.16) \quad \alpha = \Delta Y/\Delta A = 1/[1 - c(1 - t) + m]$$

So if A changes then Y changes by α times the change in A, so if A changes by one dollar then Y changes by $\alpha = 1/[1 - c(1 - t) + m]$. The denominator of the expression for a is less than unity, so the multiplier exceeds one. This is to be expected since, after the initial increase of autonomous expenditure (ΔA), any induced increase in the consumption of domestically produced goods and services will lead to a greater increase in national income (ΔY), that is, ΔY exceeds ΔA.

By inspecting the terms that define the expenditure multiplier, you can see that it is a ratio involving the marginal propensity to consume (c), the marginal tax rate (t) and the marginal propensity to import (m).

Applying the tools you learned in Chapter 7, you can observe the following:

- Other things being equal, the higher is the marginal propensity to consume, the higher is the expenditure multiplier.
- Other things being equal, the lower is the tax rate, the higher is the expenditure multiplier.
- Other things being equal, the lower is the marginal propensity to import, the higher is the expenditure multiplier.
- The opposite is the case if the marginal propensity to consume is lower, and the tax rate and marginal propensity to import are higher.

The task now is to explain the economic processes that lead to these conclusions.

We start with the essential insight that aggregate demand drives output, which generates incomes (via payments to the productive inputs). Accordingly, what is spent generates output and income in that period. The income is then available for use. There are various ways in which the income derived from the payments arising from output production can be used, namely:

- Consumption expenditure
- Saving
- Meeting tax obligations to government
- Spending on imports.

For example, workers, who are hired by firms to produce goods and services, spend part of their earned wage income on consumption. They also meet their tax obligations and may save a portion of their disposable income.

A graphical treatment

In Figure 15.4 we saw that if any of the components of autonomous aggregate expenditure change, the aggregate demand function shifts up or down, with the extent of the shift being measured by the change in the vertical intercept.

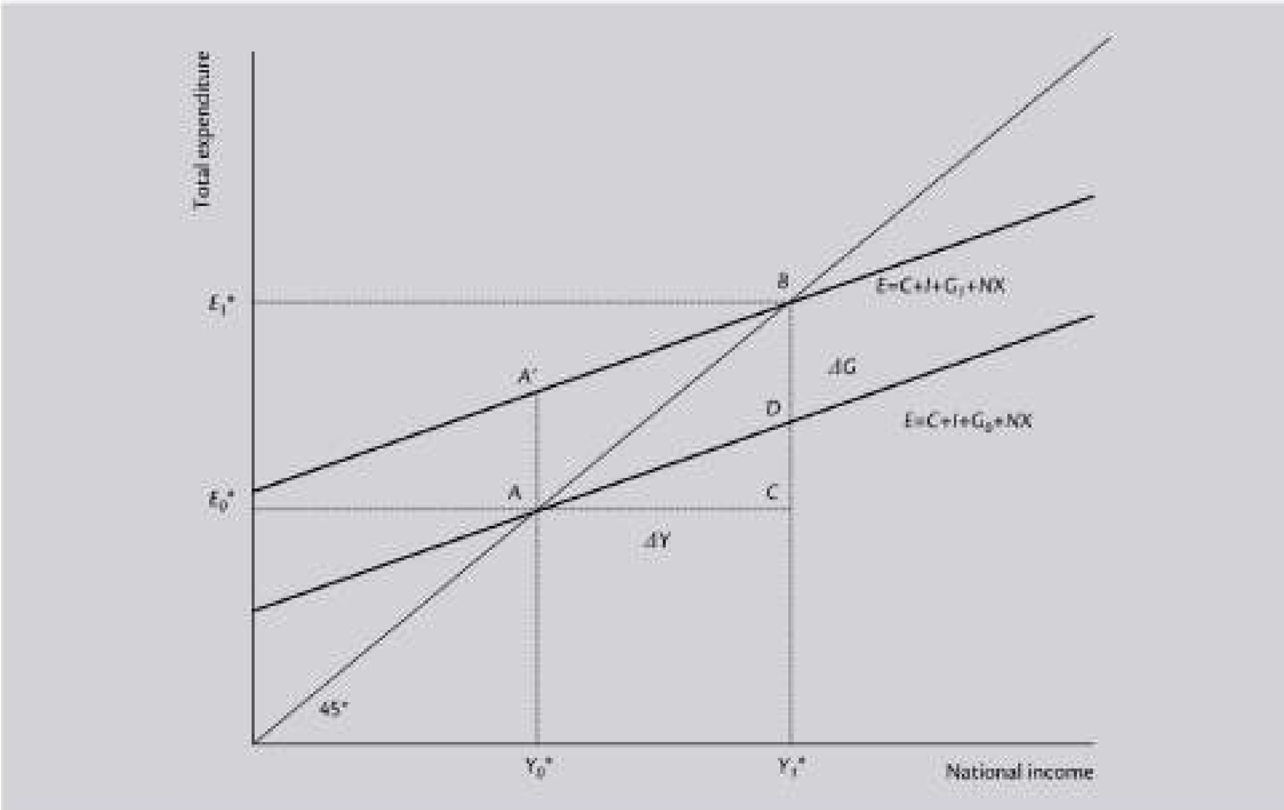
Assume that government spending rises as a result of the government being concerned that the rate of unemployment is too high. In Chapter 14 we learned that mass unemployment is always the result of deficient aggregate demand relative to the productive potential of the economy, and a simple remedy is for governments to increase total spending.

Figure 15.8 shows the change in equilibrium expenditure and income when government spending increases (ΔG). Point A is the initial level of equilibrium national income (and GDP), Y_0^*, which corresponds to aggregate expenditure of E_0^*. The aggregate demand function is given as $E = C + I + G_0 + NX$.

At this point there are no unplanned inventory changes and firms' production decisions are based upon expected aggregate demand being realised.

Now government spending increases by ΔG, which increases the aggregate demand function (where $E = C + I + G_1 + NX$) and national income increases to Y_1^*, which corresponds to aggregate expenditure of E_1^*. The new equilibrium national income is at Point B.

Figure 15.8 Impact of a change in government spending on equilibrium expenditure and income



The reason that equilibrium national income (and GDP) increase relates to the firms' revision of expected expenditure. When the government injects the new autonomous spending into the economy, aggregate spending at the current equilibrium is greater than output. The difference is the line segment AA' in Figure 15.8. This distance indicates the excess aggregate demand (relative to current GDP), and the stock of inventories would be falling. Firms would soon revise their expectations of aggregate demand upwards and start to produce more output and generate higher levels of national income.

They would continue to increase production and national income until their aggregate demand expectations were consistent with actual aggregate demand, a state which occurs at Point B (where the new aggregate demand function cuts the 45° aggregate supply line).

Note that the change in equilibrium national income (ΔY) is greater than the initial change in autonomous expenditure (ΔG). The difference between the two changes is given by the line segment CD.

How do we explain this difference?

The expenditure multiplier indicates by how much national income changes when there is a change in autonomous expenditure. The larger is the multiplier, the larger is the change in national income for a given change in autonomous expenditure.

You can see from Figure 15.8 that the total change in aggregate demand (ΔE) following a change in autonomous expenditure (in this case, ΔG) is the sum of ΔG (segment BD) and the induced consumption spending (segment DC) that follows the initial rise in national income. This is the process illustrated in Figure 15.7.

The induced consumption spending is shown as 'Induced domestic spending' in Figure 15.7. As firms react to the initial disequilibrium at Point A in Figure 15.8 (the excess aggregate demand AA') by increasing national income, households in turn, increase their consumption expenditure. But at the same time, imports are rising by $m\Delta Y$, tax revenue is rising by $t\Delta Y$ and households save a portion of each extra dollar of disposable income, $(1 - c)\Delta Y_d$.

These leakages mean that each subsequent round of induced spending is smaller than the previous one and eventually becomes zero. At that point, the economy reaches the new equilibrium at Point B in Figure 15.8.

So the total change in output and national income, ΔY is equal to the total change in aggregate expenditure, ΔE, which is equal to the initial change in autonomous spending, ΔA plus the induced consumption ΔC.

Numerical example of the expenditure multiplier at work

In our example there is an initial spending increase of $100, which might have been a government order for new public school buildings. This extra $100 in government spending leads construction firms to produce more output and increase total income payments by $100 (under the assumption in this chapter that firms are quantity adjusters). Some of this $100 is earned by construction workers in the form of wages. These workers and other income recipients then spend some of the additional income on goods and services produced locally and further afield. Refer back to Figure 15.7 to reinforce your understanding of the sequence of events.

Assume that the marginal propensity to consume is 0.75, the current tax rate is 0.20, and the marginal propensity to import is 0.20. This means that for an extra $100 of national income:

- $20 goes to tax revenue and is drained from the domestic economy.
- Disposable income thus rises by $80 and household consumption rises by $60 with the residual being additional saving of $20.
- $20 is spent on additional imports and is lost to the domestic economy.
- Total leakages from the initial $100 of extra income that is generated, namely taxes ($20) plus saving ($20) plus imports ($20), are thus $60, leaving additional consumption on **domestically produced** goods and services at $40.

The way to think of the second round expenditure injection in Figure 15.7 is to note that national income rises by $40 in response to the additional consumption spending on domestically produced goods and services, which is referred to as **induced consumption**, and then to focus on the additional leakages. After taxation is taken out, consumers determine how much they wish to spend on increased consumption.

Table 15.3 shows the process for nine rounds of additional induced domestic spending following the initial rise in government spending by $100. By the tenth round, the additional spending is too close to zero to tabulate. We deliberately use the term 'round' so as not to give the impression that the adjustment follows an orderly process across actual time. The rounds are of indeterminate length and may be irregular with respect to each other.

Each successive induced spending increase is smaller than the last because of the leakages.

Table 15.3 The expenditure multiplier process

	ΔGDP [1]	ΔTaxes [2]	ΔDisposable income [3]	ΔConsumption [4]	ΔSaving [5]	ΔImports [6]	ΔTotal leakages [7]
Round 1	100.0	20.0	80.0	60.0	20.0	20.0	60.0
Round 2	40.0	8.0	32.0	24.0	8.0	8.0	24.0
Round 3	16.0	3.2	12.8	9.6	3.2	3.2	9.6
Round 4	6.4	1.3	5.1	3.8	1.3	1.3	3.8
Round 5	2.6	0.5	2.0	1.5	0.5	0.5	1.5
Round 6	1.0	0.2	0.8	0.6	0.2	0.2	0.6
Round 7	0.4	0.1	0.3	0.2	0.1	0.1	0.2
Round 8	0.2	0.0	0.1	0.1	0.0	0.0	0.1
Round 9	0.1	0.0	0.1	0.0	0.0	0.0	0.0
Total change	166.6	33.3	133.3	100.0	33.3	33.3	100.0

Note: Figures have been rounded to 1 decimal place.

The initial spending 'fans out' or spreads throughout the entire economy. The initial spending therefore **multiplies** into a much larger increase in spending.

Note that the **leakages** are getting smaller with each spending period. In the bottom row of Table 15.3, we see that the sum of all the period-by-period changes in GDP equals $166.6. The total increase in tax revenue is $33.3. The total increase in disposable income is thus $133.3, which leads to total induced consumption of $100 and the total increase in saving of $33.3. Total imports rise by $33.3.

The total additional leakages from the expenditure system – taxes, saving and imports – sum to $100.00 at the end of the adjustment period, and this is the amount of the total initial injection in autonomous expenditure. At the point when the additional leakages equal the additional injections, the system is at rest and the multiplied impact of the initial injection in autonomous expenditure is complete.

Think back to our algebraic definition of national income equilibrium in Equation (15.15) and the related expression for the expenditure multiplier, Equation (15.16). If we substitute the assumed values in this example into the multiplier formula we get:

$$\text{Multiplier} = \Delta Y/\Delta G = 1/[1 - c(1 - t) + m] = 1/[1 - 0.75(1 - 0.20) + 0.20] = 1.666$$

The multiplier is the total change in GDP for a one dollar initial increase in aggregate demand, which we have calculated to be 1.666. This means that if autonomous expenditure (for example, government spending) rose by $100, the total change in GDP, after the economy adjusts to the higher production and income levels, would be $166.7 (rounded), which matches the analysis in Table 15.3.

We have thus also derived a condition for macroeconomic equilibrium following a disturbance in one of the exogenous expenditure components: **the economy restores equilibrium when total leakages exactly equal total injections**.

Changes in the magnitude of the expenditure multiplier

Equation (15.16) defined the expenditure multiplier as $\Delta Y/\Delta G = 1/[1 - c(1 - t) + m]$, which means that its size depends on the marginal propensity to consume (c), the tax rate (t) and the marginal propensity to import (m).

The following conclusions can be drawn at this stage:

- The multiplier is larger(smaller), the larger(smaller) is the marginal propensity to consume (c).
- The multiplier is larger(smaller), the smaller(larger) is the marginal propensity to import (m). The more open the economy is to trade, the lower is the multiplier.
- The multiplier is larger(smaller), the smaller(larger) is the tax rate (t).

A higher marginal propensity to consume means that each successive round of induced consumption spending is larger, other things being equal. Given that the marginal propensity to save (s) is just $(1 - c)$, the multiplier is higher when the marginal propensity to save is lower.

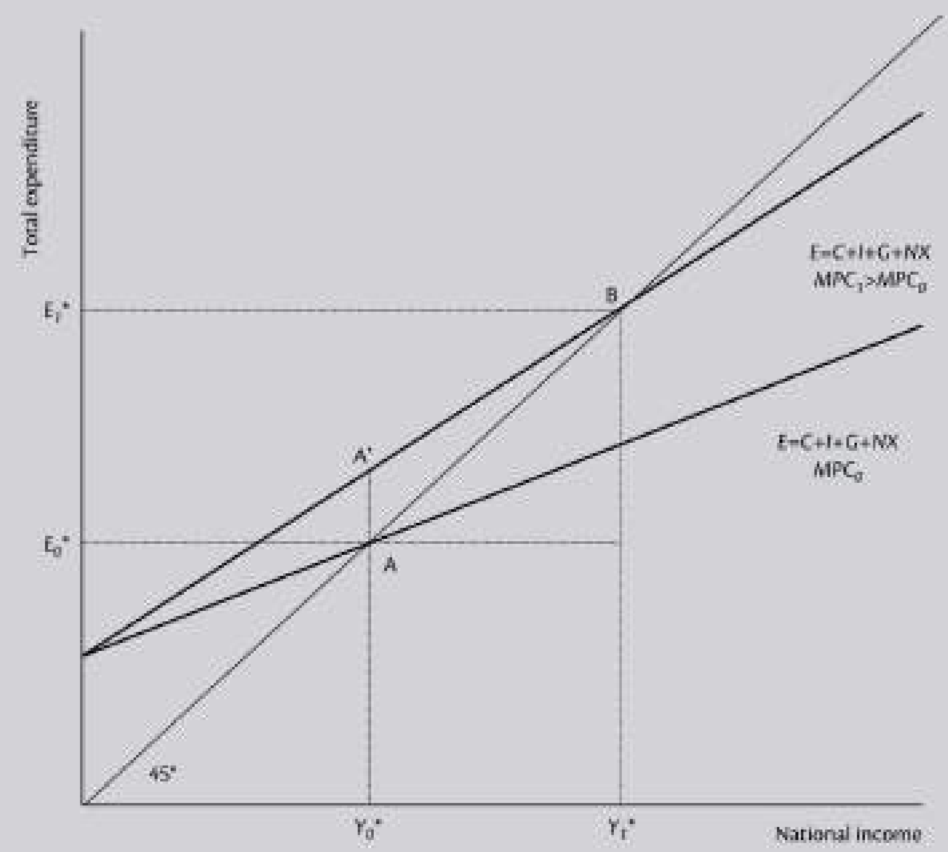
In general, the lower are the leakages (taxes, saving and imports) from the expenditure system, the higher will be the multiplier. This is because the lower the leakage from each spending round, the larger is the induced consumption.

Earlier in the chapter, we learned that the slope of the aggregate demand function (see Figure 15.4) was $[c(1 - t) - m]$, which tells us that the change in aggregate spending for a given change in national income is larger, the larger is the marginal propensity to consume (c) and the lower is the tax rate (t) and the marginal propensity to import (m).

Consider an increase in the marginal propensity to consume (c). Figure 15.9 depicts this case. The initial aggregate demand function (for MPC_0) is associated with a national income equilibrium at Point A where $E_0{}^*$ generates production and national income of $Y_0{}^*$.

When the marginal propensity to consume rises to MPC_1, the aggregate demand function pivots upwards at the vertical intercept. At the current equilibrium income level, $Y_0{}^*$, households choose to spend an increased proportion of each dollar of disposable income on consumption. The initial change in aggregate demand (E) is measured by the distance A to A'.

Figure 15.9 Impact of a change in the marginal propensity to consume on equilibrium expenditure and income



The economy responds to the increased consumption spending by increasing production and national income. The higher multiplier (as a result of the higher MPC) then drives national income up further and the economy reaches a new equilibrium at Point B. At that point the economy comes to rest again.

The following conclusions can be drawn at this stage:

- The slope of the aggregate demand function is steeper(shallower) the larger(smaller) is the marginal propensity to consume (c). This means that for a given flow of autonomous spending, aggregate demand and national income will be higher if the marginal propensity to consume rises, and vice versa.
- The slope of the aggregate demand function is steeper(shallower) the smaller(larger) is the marginal propensity to import (m). This means that for a given flow of autonomous spending, aggregate demand and national income will be higher if the marginal propensity to import falls, and vice versa.
- The slope of the aggregate demand function is steeper(shallower) the smaller(larger) is the tax rate (t). This means that for a given flow of autonomous spending, aggregate demand and national income will be higher if the tax rate falls, and vice versa.

A final point about the multiplier

We have described the multiplier as a process through which an increase of autonomous expenditure raises induced consumption and thus income through a series of steps to reach a new higher level of expenditure and income. We have defined this as a point of equilibrium in the sense that once we reach this point, we are in a state of rest because no one has an incentive to change their spending further. We must be careful in our interpretation of the process, however. We must not conclude that spending and income are unequal while the multiplier process is unfolding. Recall from our discussion of the D and Z curves in Chapter 13 that the intersection of the two occurs

TRY IT YOURSELF

[illegible] shows the impact of varying the parameters that determine the size of the multiplier (c, m and t) on national income for a given injection of autonomous expenditure. You might like to simulate multiple changes in the parameters (for example, a rising MPC and a rising tax rate) to gain a greater understanding of how these influences interact.

Table 15.4 Simulating changes in the multiplier components

Varying the marginal propensity to consume			
Marginal propensity to consume (c)	0.6	0.7	0.8
Marginal propensity to import (m)	0.2	0.2	0.2
Tax rate	0.2	0.2	0.2
Multiplier	1.39	1.56	1.79
Autonomous spending	100	100	100
National income	139	156	179
Varying the marginal propensity to import			
Marginal propensity to consume (c)	0.8	0.8	0.8
Marginal propensity to import (m)	0.1	0.2	0.3
Tax rate	0.2	0.2	0.2
Multiplier	2.17	1.79	1.52
Autonomous spending	100	100	100
National income	217	179	152
Varying the tax rate			
Marginal propensity to consume (c)	0.8	0.8	0.8
Marginal propensity to import (m)	0.2	0.2	0.2
Tax rate	0.1	0.2	0.3
Multiplier	2.08	1.79	1.56
Autonomous spending	100	100	100
National income	208	179	156

at the point of effective demand. This point in turn is consistent with the expectations of firms, which are hiring the quantity of labour they think they need to produce the amount of output they expect to sell. Firms do not know in advance that their expectations will be met. Still, it is their expectation of sales that determines the point of effective demand, and it is an equilibrium in the same sense: no one has an incentive to do anything differently.

In the context of this chapter, one month's point of effective demand can be larger than the previous month's. Let us say that firms have become more optimistic and so have increased their spending on plant and equipment, raising investment. This chapter has taught us that higher investment spending will induce more consumption through the multiplier. This is because more employment in the investment sector (producing more plant and equipment) generates more wages and hence more consumption by workers. Over the course of the month, stores will experience higher than expected sales, which they could not have foreseen because their expectations for the month did not take into account the new, higher, employment.

The unexpected extra sales will run down their stock of goods on the shelves. We categorise this as a decline of inventories, which is treated in the National Income and Product Accounts as negative investment. At the aggregate level, our accounting will show that the reduction of inventories offsets (at least to some extent) the increased spending on plant and equipment.

However, as stores draw down their inventories, they will place orders with producers of consumer goods to restock the shelves. This in turn will signal to the producers that they need to increase production. At first they can do this by increasing overtime work. However, as they come to expect the new, higher sales to persist, they will hire more workers to meet the higher demand for consumption goods.

You can see how the multiplier process will work its way through the economy as a whole. In this particular case it first increases production of plant and equipment and then production of consumer goods to restock shelves, and then on to an entire array of consumer goods and services, which can then trigger more investment to increase capacity. In terms of our D-Z curve analysis, the point of effective demand continues to shift out to the right as firms raise their expectations of sales and hence increase employment. But at each point during the multiplier process, the employment decision is consistent with a point of effective demand, that is, equilibrium.

Of course, this process can work in reverse should investment decline. Through the multiplier process firms will find their expectations disappointed (stores experience falling sales and thus rising inventories) and will cut back employment and output. But again, each employment decision will be consistent with a point of effective demand.

This is how changing expectations about the future affect the level of employment today.

Conclusion

In this chapter we provided a detailed examination of aggregate demand and aggregate supply. We have distinguished between the macroeconomic identities (those behind the National Income and Product Accounts) and behavioural equations that allow us to say something about causation.

We considered three important propositions about the macroeconomy:

1. Total spending drives total output (GDP) and employment in the economy.
2. There is no guarantee that equilibrium output will be associated with full employment.
3. A change in autonomous expenditure, such as investment or government expenditure, will lead to a multiplied (larger) change in GDP (national income).

In Chapter 16 we will complete the picture by developing a more detailed aggregate supply framework. We can do that by using the National Accounting concept of GDP at constant rather than current prices as our measure of economic activity.

References

ABS (Australian Bureau of Statistics) (2017) "Expenditure Chain Volume Measures", *Australian National Accounts: National Income, Expenditure and Product*, September (Cat. 5206.0).

Keynes, J.M. (1936) *The General Theory of Employment, Interest, and Money*, London: Macmillan, 1957 Reprint.

OECD (2017) "Household Final Consumption Expenditure, Percentage of GDP", *OECD* National Accounts Statistics (database). Available at: http://www.oecd-ilibrary.org/economics/data/oecd-national-accounts-statistics/national-accounts-at-a-glance_data-00369-en#, accessed on 15 April 2017.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

16 AGGREGATE SUPPLY

Chapter Outline

Learning Objectives

- Understand the mark-up pricing model and its underlying assumptions.
- Explain why the pricing model is consistent with firms acting as quantity adjusters.
- Recognise that labour productivity can be procyclical.

16.1 Introduction

In Chapter 15, our theory of expenditure and income determination linked aggregate spending to the generation of income. The focus on the demand drivers of aggregate income and output abstracted from any spending impacts on the price level and assumed that most firms in the economy were rather passive. They simply responded to an increase in nominal spending by increasing real output up to the full capacity level in the economy. In doing so, we ignored the complexity of the supply side. We also abstracted from what might happen after the economy reached its full capacity level.

In this chapter, we seek to explore how the economy responds to an increase in nominal aggregate spending. The simple reverse L-shaped supply curve we discussed in Chapter 12 *Mr. Keynes and the 'Classics'* is too simple to represent real world behaviour. More generally, each firm has three possible responses:

1. To behave as a quantity adjuster and increase output.
2. To behave as a price adjuster and increase prices for its output.
3. A combination of both forms of adjustment.

In Chapter 15, we assumed that firms respond by increasing their output (response 1) when confronted with higher demand for their products (quantity adjustment). However, in the real world, the other two possibilities might also be observed. In this chapter, we consider the circumstances under which firms might deploy these different responses.

To capture price setting behaviour, we develop the widely used model of **mark-up pricing**, where firms have some market power and set prices to achieve a target profit margin over unit costs. This model provides a rationale for the claim that over a normal range of capacity utilisation and output, the price level is more or less constant. This means that, as a first approximation, treating firms as quantity adjusters in response to changing levels of total expenditure is a reasonable assumption.

However, we recognise that once full capacity is reached, firms can no longer increase production and therefore will respond to rising nominal expenditure (demand for their goods and services) using price rationing strategies (response 2). Further, we also recognise that different sectors in the economy will reach full capacity at different times and so some price increases might begin to be observed before the overall economy is at full capacity (response 3).

This was Keynes' view in Chapter 20 of *The General Theory*, which presented the 'employment function'. He argued that the elasticity of employment (and production) with respect to an increase of demand varies across industries. When demand rises, some combination of output and price increases will absorb the increased demand, with the proportions varying across industries up to the point of economy-wide full employment. Once that is reached, only prices can rise because firms cannot find more resources to produce more.

The theory of **aggregate supply** thus seeks to explain the factors that impact on a firm's decision to supply output in response to expected aggregate spending. It must therefore incorporate the three responses outlined above. The theory we develop in this chapter will therefore complete the demand side model that we developed in Chapter 15 to allow us to present a comprehensive theory of the determination of the output level, the price level and total employment.

16.2 Some Important Concepts

Schedules and functions

In this chapter we will consider aggregate supply schedules.

The terms 'schedule' and 'function' are used interchangeably in the economics literature. We prefer to use function to depict a relationship between variables, such as spending and income.

REMINDER BOX

In Chapter 7 *Methods, Tools and Techniques* we introduced the essential analytical and introductory techniques that students should learn in order to grasp macroeconomics.

As a reminder, economic models use schedules or curves to depict behaviour, which can either be *ex ante* (prior to action and reflecting planned or desired action by households, firms, government and so on); or *ex post* (representing actual outcomes that are the results of action).

In the simplest macroeconomic model of expenditure, income and employment we encounter an aggregate demand schedule and an aggregate supply schedule. These schedules depict *ex ante* behaviour and tell us what we expect the outcomes will be, given other conditions in the economy.

The employment-output function

First we consider how total employment in the economy is generated so that we can understand what determines **unit labour costs**. To develop a theory of employment, that is, explain its level and movement over time, in relation to a monetary economy operating under capitalist conditions, we need to develop an understanding of how employment is related to output determination. This relationship is also important because per-unit labour costs (total labour costs divided by total output) underpin the pricing of output via a **price mark-up**. In this context, we develop the concept of the **employment-output function**, which shows how much labour is required to produce a given volume of output.

Given the output that the firm plans to produce to meet expected demand, employment will be determined by the productivity of labour. We shall argue that production decisions are typically made in an environment of stable wage rates and capital-labour ratios. The capital-labour ratio depicts the combination of productive capital (machines, equipment and so on) and labour that defines the current productive technology.

For example, an excavation firm might provide a hand shovel to each worker engaged in digging foundations for a new building. This would be a low capital-labour ratio production technology. Sometimes this is referred to as a labour-intensive technique. Alternatively, it could use mechanical digging equipment and employ fewer workers to produce the same output. In this instance, the production process would employ higher capital-labour ratio techniques, sometimes referred to as capital-intensive production.

We can write the **employment-output function** as:

$$Y = \gamma N \tag{16.1}$$

where N is the total number of workers employed, γ is the rate of labour productivity, and Y is planned production (based on expected spending).

What is labour productivity? Labour productivity is defined as physical output per unit of labour input per period of time. So, we could solve Equation 16.1 for γ to get Y/N, which is the algebraic equivalent of our definition.

The higher is labour productivity (γ) the less employment is required to produce a unit of output for the given production technique (implicit in γ).

Factors which influence γ include: technology (is it best practice, capital or labour intensive – as in our excavation firm example); worker skill and motivation; and management skill and business organisation. In the public arena, discussions about slowing productivity growth often focus unduly on the worker with claims such as poor motivation and skill gaps. Rarely is management skill (or failings thereof) the focus of enquiry despite evidence that poor management decision making is a cause of slow productivity growth. For instance, failure to invest in the latest technology will hamper growth in labour productivity, as demonstrated in Box 16.1.

BOX 16.1 THE PERILS OF NEGLECTING INNOVATION

In the 1950s, the large American steel companies fell behind their Japanese and European competitors because they failed to scrap the old 'open hearth furnaces' and invest in the latest blast furnaces. In the 1960s, they also failed to convert to continuous casting processes, which delivered superior productivity.

A more recent example can be found in the airline industry. In the late 1970s the Australian airline Qantas dominated the international travel market for Australians, carrying around 42 per cent of Australian travellers abroad. By 2012, this proportion had dropped to 18 per cent as competition from airlines such as Emirates and Singapore Airlines had cut into its market share. There are many reasons for this decline in market share, but one of the major explanations was that Qantas management made poor decisions with respect to its fleet upgrades. It refused to invest in the latest jets which were more fuel efficient and hence could operate at lower cost.

If γ is stable in the short run (within the current investment cycle), then once the firm decides on the level of output to produce to satisfy expected demand, it simultaneously knows how many workers must be employed. As an example, if it takes ten workers to produce 1,000 units of output per day, then daily labour productivity would be 100 units per worker. Accordingly, if the firm anticipated an increase in output to say 1,500 units per day, it would require an additional five workers to ensure it could supply the new higher level of output.

Figure 16.1 presents two different employment-output functions for the economy. Each is associated with a constant but different value of γ, and the functions are positively sloped straight lines.

If a firm's expected aggregate demand was 1,200 units in the current production period, say per week, then given the state of technology (represented by γ) it would employ 600 workers each week if $\gamma = 2$ (lower productivity) and 400 workers if $\gamma = 3$ (high productivity).

Firms produce based on expected aggregate spending and once all the sectors have made their spending decisions (that is, once aggregate demand is actually realised), the firms discover whether their expectations were accurate or not. In other words, they find out whether they have overproduced, underproduced or produced the right amount, only after spending has occurred.

Money wages

Productivity is an important component of the cost of producing each unit of output (unit cost). Another major component of unit cost are labour costs, which is influenced by the prevailing wage rates.

We assume that money wage rates are exogenous in the short run. This is not the same as assuming that money wage rates never change. It merely says that, in terms of the parameters of our aggregate supply model (that is, the different influences that we consider will impact on aggregate supply), the money wage rate will be assumed to be invariant in the short run.

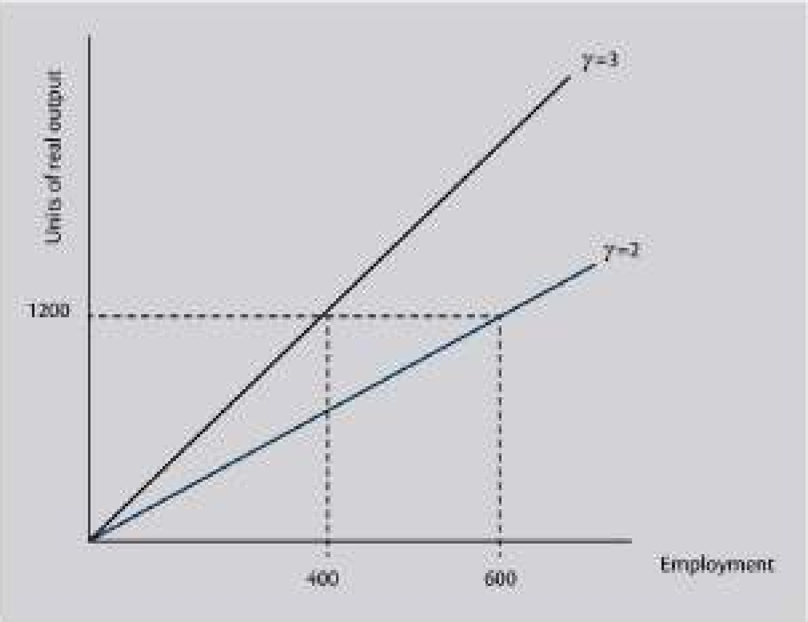
Before we discuss the possible factors which make this a reasonable assumption, we must clarify some often confused concepts relating to wages (see also Chapter 12).

First, let us distinguish between the **money wage** rate and the **real wage** rate.

The money wage rate is determined in the labour market and is the amount in nominal (current dollar) terms that the workers receive per hour (or some other period) when they sell their labour power to the capitalist business firms or other employers (for example, government). The actual money wage at any point in time is the outcome of negotiated agreements between employers and workers, either at a decentralised level or through sector or economy-wide negotiations between peak employer groups and trade unions. These money wage bargains are negotiated in the context of prevailing government policy with respect to minimum (or living) wages. In some nations, such as Australia, there has been a history of wage setting tribunals (courts) which has led to the practice of industrial relations, in general, and the determination of wages, in particular, becoming a specialised judicial process, thereby reflecting the adversarial nature of relations between workers and capital.

The money wage outcome at any point in time is heavily dependent on

Figure 16.1 The employment-output function



the bargaining strengths of the parties involved. Wage changes occur at infrequent intervals and condition the behaviour of the parties concerned for the ensuing economic period (sometimes months, usually years). It is this infrequent nature of wage setting via institutional structures (such as employer and union negotiations) and the implied contractual nature of the wage relationship existing between employers and workers over some future period that are used to justify the assumption that money wages are exogenous and fixed in the short run for the purpose of developing an explanation of aggregate supply.

What is the basis of the money wage inflexibility assumption? First, negotiations over money wages typically occur at infrequent intervals, as noted above. Second, there is strong evidence that workers resist cuts in money wages and firms generally prefer not to push for such cuts. This is because the structure of wages across all occupations and sectors represents an indicator of social status. Workers are very aware of wage relativities and are loath to agree to a reduction in their relative rank given that it signals a reduction in their social status. One group of workers will form the view that in an economic downturn, if they accept lower money wages and other workers resist the reductions, then their relative position is compromised and there is no certainty that the relativity will be restored when economic conditions improve. Only in extraordinary circumstances relating to the imminent collapse of the enterprise in which they are employed and the existence of very high levels of unemployment have we observed workers agreeing to money wage reductions.

Third, the downward rigidity of money wages is also the result of employer preferences. Even when the unemployment rate approaches double digits (a rate considered high by historical standards), the absolute number of workers not in employment relative to those who retain their jobs is small. As such, employers are reluctant to risk jeopardising convivial industrial relations with most workers to possibly improve the employment prospects of a small proportion of employed workers. The costs of being a capricious employer in a downturn can come back to haunt a firm when times improve and they find that workers prefer to take jobs in other firms, perhaps their competitors.

We consider these issues in more detail in Chapter 17 *Unemployment and Inflation*.

The real wage rate is the **purchasing power equivalent** of the money wage rate, that is, how much output the worker can buy with a dollar of their money (or nominal) wage. The real wage is calculated by deflating the money wage by a price index. We learned how deflators are constructed and are used to convert current price variables into constant price (real) variables in Chapter 4.

The choice of deflator depends on the context. The real wage from the perspective of the worker would be the money wage expressed in terms of consumption good equivalents. So, we would consider this rate to be the money wage rate divided by a measure of consumer prices (such as the CPI).

From the employer's perspective, the real (product) wage is more accurately measured by the money wage paid to workers divided by the specific price the firm receives for its output, which is a narrower concept than the real wage considered from the perspective of the worker.

It is unlikely that the path of prices received for the output of any particular firm over time would mirror the path of consumer prices as a whole. This must always be taken into account when we distinguish impacts on individual firms versus impacts on the economy as a whole. Workers will want their nominal wages to grow at least as fast as consumer prices grow so that their real wages do not decline. On the other hand, many firms will be unable to increase their prices at this rate, meaning that the real wage they face would be rising if they were to agree to raise the nominal wages they pay at the rate that consumer prices are growing.

Importantly, the real wage is not determined in the labour market and can only be influenced by the workers inasmuch as they can influence the money wage rate outcome. This is because **the real wage is a ratio of two prices**, the money wage (determined in the labour market) and the consumer price level (determined in the goods and services market and influenced by the price setting behaviour of firms). The two prices that form the real wage rate are determined by different forces in different markets in the economy.

As we will see, prices are largely set by business firms in the goods and services (product) market according to desired mark-ups on cost. Prices are not fixed by workers.

Economists and others often draw on Classical employment theory (see Chapter 12) and argue that workers should cut their real wages to improve the employment prospects of the unemployed. However, this policy suggestion is without merit. Even if the proposition were based on a causal understanding of how mass unemployment occurs, there are several preliminary, but critical questions that such proposals fail to answer:

- How can workers achieve a cut in their real wage when they can only influence the money wage outcome?
- How might a money wage change influence price changes? In particular, a money wage cut may lead to price cuts due to the fall in the costs of production, and thus leave the real wage unchanged.

These initial queries stand quite apart from the dispute among economists as to whether a real wage cut would influence employment growth independently of changes in effective demand. As we saw in Chapter 12, employment is determined by total output rather than the real wage.

16.3 Price Determination

Clearly, a firm seeks to generate a profit over and above the costs of production. How does it go about setting the price that it will accept for its output to achieve its ambitions for profit?

Firms are assumed to operate in a **non-competitive** economy. You may have considered the case of perfect competition in a microeconomics course where firms are assumed to have no price setting discretion because the market is so large and firms are assumed to be so small. That environment does not arise in the real world.

We are thus introducing oligopoly as a basic assumption rather than following the orthodox practice of using perfect competition as the benchmark. By oligopoly, we mean a product market in which there are a small number of sellers. Hence, we assume that firms are **price setters rather than price takers.**

Firms are assumed to fix their prices as a mark-up over costs. Economists are divided about the determinants of the mark-up and the costs that are considered relevant in the pricing decision by firms. Further, debate remains as to whether the mark-up is invariant to the state of demand.

However, the use of the mark-up as a basic description of firm behaviour in the real world is difficult to dispute. In the real world, firms typically have discretionary price setting power and seek a rate of return on the capital employed, which necessitates that they generate a profit margin over their total costs of production.

The total price per unit sold must therefore cover its (variable) costs of production per unit of output, such as labour and raw material costs, plus the profit margin. The profit margin is designed to cover overheads and other fixed costs plus net profit.

Firms are thus assumed to employ a mark-up pricing model such that:

$$P = (1 + m)[W/\gamma] \tag{16.2}$$

where P is the price of output, m is the mark-up on per-unit labour costs, W is the money wage per hour and γ is labour productivity per hour. At this stage we abstract from raw material costs. Thus γ is defined as the units of output per unit of labour input per hour.

If $\gamma = 0.5$ then two labour hours are required to produce one unit of output. If the money wage (W) was \$5 per hour, then the **unit labour costs** (that is, labour cost per unit of output) would be \$10.

As noted, the mark-up (m) is set to provide a surplus above the direct unit labour costs to account for fixed (overhead) labour and other fixed costs, including interest payments on loans, in addition to a provision for profits (return on equity). This is a gross profit measure, as the firm must also cover taxes and other business expenses (accounting services, advertising, and legal expenses) out of the mark-up. The amount of profit desired is related in part to the amount of investment that the firms plan to undertake because retained earnings are an important source of internal finance that the firm draws on to reduce its exposure to the higher costs of externally funding new projects.

In the short run, the price will be rigid with the firm supplying output according to demand. Price changes would occur when there were changes in the money wage rate or other variable costs, the mark-up (margin), or trend labour productivity. Trend labour productivity is used here to differentiate it from the cyclical swings that occur in labour productivity, which we consider in Section 16.6.

The mark-up or margin (m) reflects the **market power** of the firm. The higher the market power, the higher will be the margin. Thus, in more competitive sectors, the margin will tend to be lower than in less competitive sectors. Changes in competitiveness of a sector will, over time, lead to changes in the size of the mark-up.

If, in our example, the mark-up (m) is set at 40 per cent, then the firm will price its output at \$14 per unit (\$10 multiplied by 1.40).

The features of this approach to price determination are as follows:

1. Prices are unambiguously a function of costs.
2. Firms use their price setting discretion to generate a monetary surplus above average variable costs. This monetary surplus is designed to cover profits. In the short run, profits are influenced by the ability of firms to realise the mark-up on their unit costs. Factors which squeeze the mark-up (down to say 30 per cent) will accordingly also squeeze profits per unit of output.
3. The volume of profits (as distinct from the per-unit profit) depends on the size of the mark-up, which influences profit per unit of output, and the actual volume of output sold in any period. The latter is determined by the state of aggregate demand in the economy and, as we saw in Chapter 15, is determined by the level of household consumption expenditure, private investment expenditure, net exports and government spending.
4. Usually mark-up theories assume that the immediate impact of changes in demand on the mark-up, and hence prices, is small. For the planning period ahead, firms calculate their costs and desired profits on the basis of an expected level of output, which they believe they can sell. Deviations in this expected level of demand promote output changes rather than price changes. For example, it is expensive to alter prices once catalogues are advertised, although the rising use of on-line shopping sites has somewhat reduced this form of inflexibility. Firms also desire to be seen as reliable suppliers at stated prices.
5. The mark-up impacts directly on the real wage that workers receive. Assume (for simplicity) that total marked-up costs only include wage costs. Total wage costs are the product of the money wage rate W and the number of workers employed N, that is, WN.

In this case, a simplified price mark-up model would be:

$$(16.3) \qquad P = (1 + m)W/\gamma = (1 + m)WN/Y$$

where all the terms are as defined previously, in particular average labour productivity, $\gamma = Y/N$. WN/Y are wage costs per unit of output, in other words per-unit labour costs, which we defined above. We can rewrite this as:

$$(16.4) \qquad Y/(1 + m) = WN/P$$

and further rearrangement yields:

$$(16.5) \qquad W/P = (Y/N)/(1 + m) = \gamma/(1 + m)$$

which shows that the real wage (W/P) is dependent on the average productivity of labour (Y/N) and the size of the mark-up. The larger the mark-up (m), other things being equal, the lower is the real wage.

16.4 The Aggregate Supply Function (AS)

Before we consider complicating factors, such as changes in productivity and competitiveness, it is useful to consider what the price determination rule means for the shape of the aggregate supply function.

If we assume that m, W and γ are constant in the short run then the aggregate supply curve would be a horizontal line in the graph of price against real income up to some full capacity utilisation point (Y^*). Economists sometimes refer to a horizontal line in this context as being perfectly elastic. Firms in aggregate will supply as much output (goods and services) as they expect will be demanded at the current price level, set according to the mark-up rule described above.

So far we have assumed that labour costs are the only variable cost. If labour productivity is constant and other direct (variable) production costs, such as raw materials, are constant per unit of output, the aggregate supply curve will be elastic at the (constant) price associated with marking up per-unit production costs. Also, the labour demand curve at both the firm and aggregate level will be elastic at the going money wage, subject to the level of aggregate demand.

Figure 16.2 shows the way in which the price set by all firms (P_0) at a point in time is distributed as incomes. Here the current level of output which is sold is Y_0. The price P_0 is based on a mark-up on total unit variable costs which covers fixed costs (including labour overheads) and an allowance for profit.

Total spending in the overall economy is the area defined by $P_0 \times Y_0$ and the distribution of that level of output as income is shown by the areas below the price line. Fixed costs are represented by the rectangle A, whereas rectangle B represents net profit.

Firms produce a given level of output according to their expectations of total spending in the economy. For simplicity, we assume that production is in line with sales, Y_0. In fact, the total output sold may be less than firms expected and hence less than they produced. Variable costs, namely labour and raw material costs, which are incurred in producing the unsold output must still be met. This unsold output will add to the stock of inventories of final output. Conversely, if sales are higher than expected, then the stock of inventories will be reduced. In that sense, the net profits generated may be below or above the level that the firms aimed to achieve at the beginning of the production period. Firms may plan to increase profits in future by raising the mark-up.

The horizontal segment in Figure 16.2 has been explained by the price mark-up rule and the assumption of constant unit costs. But why does it become vertical after full employment?

Figure 16.2 Output, sales and national income

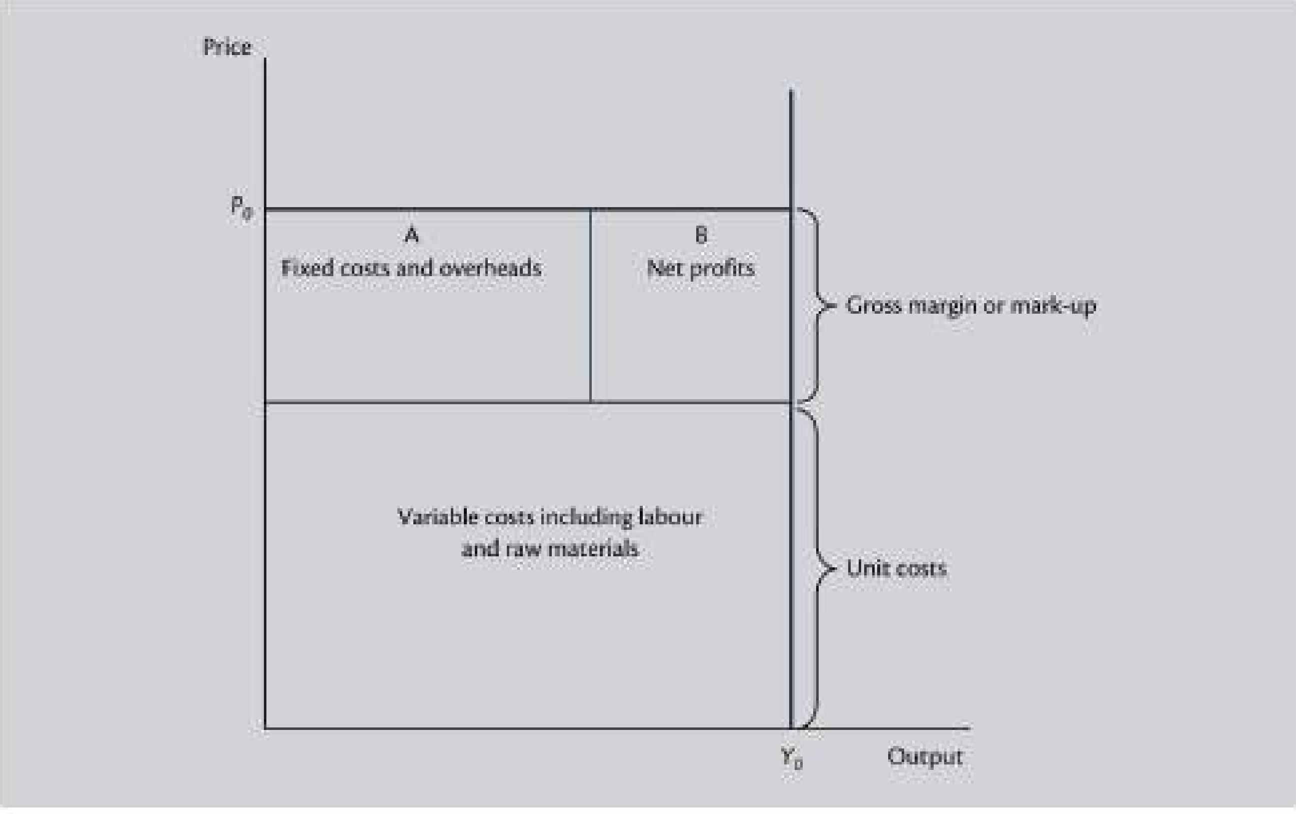


Figure 16.3 is similar to Figure 16.2 but adds the full capacity utilisation level of real output (Y^*) to derive the **general aggregate supply function (AS)**. This AS function is sometimes referred to as a reverse L shape for obvious reasons.

After the point of full employment, the economy exhausts its capacity to expand short-run output due to shortages of labour and capital equipment. Firms will be trying to outbid each other for the already fully employed labour resources and in doing so would drive money wages up. Also there is the possibility of prices of raw materials being driven up by the high levels of demand. We will return to this possibility later in this chapter.

Under normal circumstances, the economy will rarely approach the output level (Y^*), which means that for normal utilisation rates the economy faces constant costs.

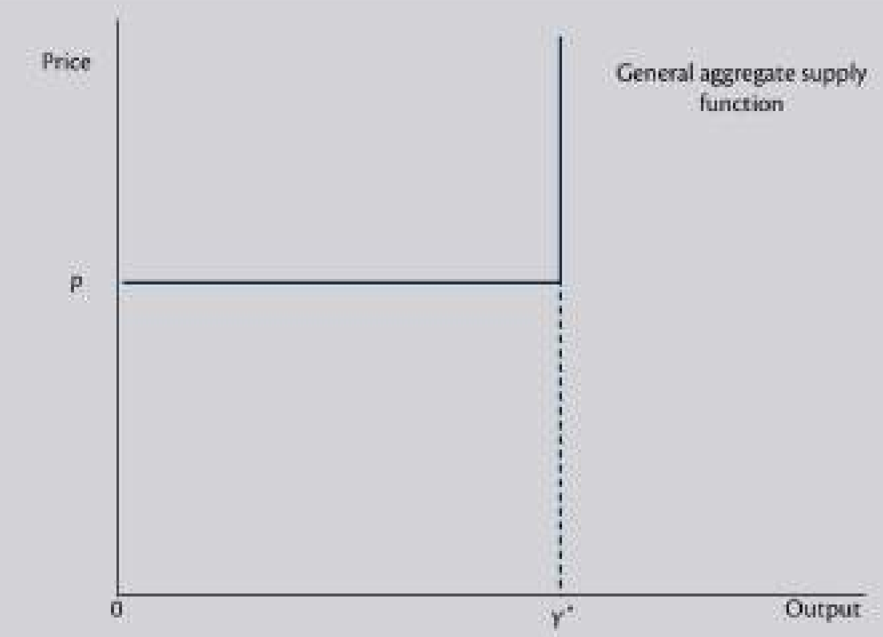
There is some debate about when the rising costs might be encountered given that all firms are unlikely to hit full capacity simultaneously. The **reverse L shape** simplifies the analysis somewhat by assuming that the capacity constraint is reached by all firms at the same time. Bottlenecks in production are likely to occur in some sectors before others and so cost pressures will begin to mount before overall full capacity output. This could be captured in Figure 16.3 by some curvature near Y^*, thus eliminating the right angle as prices begin to rise before reaching Y^* (full capacity). We consider this issue in more detail in Chapter 17.

The theory of production

The theory of production that we have presented here, which underpins the model of pricing, is based on several stylised facts from the real world.

- The capacity of firms to substitute one input (say, labour) for another (say, capital) in the production process is limited. In the real world, a typical firm employs a number of machines and types of equipment, which have more or less fixed labour requirements. Then idle machines typically accompany idle workers when the economy goes into a downturn.
- Economies are rarely at full employment so that the existing capital stock is rarely fully utilised.

Figure 16.3 The general aggregate supply function (AS)



Take the simple example of a cleaning firm which uses brooms as its principal technology. It services a contract to sweep rooms in office blocks each day. It is hard to imagine two workers pushing one broom or one worker pushing two brooms. To start production, the firm needs to combine its productive inputs in a fixed ratio (in this case, one to one).

If the firm gained contracts for more office cleaning which exceeded the capacity of one cleaner, then it would have to (given the technology being used) add another broom for the second worker to use and so on. So, the productive inputs are added in fixed proportions defined by the technology being used.

It doesn't violate reality too much to simplify this stylised fact by assuming what economists refer to as fixed input coefficients technology. Consequently, we reject the Law of Diminishing Returns, which we outlined in Chapter 11, as being an accurate representation of conditions of production in the macroeconomy. Under this so-called law, the stock of capital is always assumed to be fully utilised. Falling marginal productivity occurs when more labour is employed with the fixed stock of capital. Thus, in contrast to the fixed input coefficients technology described above, the Law of Diminishing Returns relies on the capital-labour ratio declining in the short run when employment increases, so that each worker has less capital.

Thus, the neoclassical assertion that capital, in the form of specific plant and equipment, is fixed in the short run (see Chapter 11) confuses the distinction between the stock of capital in value terms (its monetary worth), and the flow of services that the stock produces, as revealed by the rate of capacity utilisation.

The neoclassical production function analysis, which is standard in most textbooks, assumes that in the short run, with all other productive inputs (capital, land and so on) fixed, output will increase at a decreasing rate as more hours of employment are used by firms. However, in the real world, the actual relationship between changes in labour hours and changes in output may not exhibit diminishing returns because other productive inputs may vary in the same proportion as the labour input.

The neoclassical production function enables neoclassical economists to postulate increasing marginal costs as output increases (that is, costs increase faster as more output is produced). In turn, this leads to an inverse relationship between labour demand and the real wage. These relationships derive from the assumption that firms produce and employ labour such that their profits are maximised at a given price and money wage.

The validity of this 'Law of Diminishing Returns' has been the subject of considerable controversy. In essence, it is a theoretical construct; an unproven assertion. No conclusive empirical evidence has ever been assembled to substantiate this 'Law' as a reasonable generalisation of production relationships in modern monetary economies.

On the contrary, there is a mass of empirical evidence available, derived from actual studies of business firms, to support the view that costs of production are constant in the relevant or normal range of output and that the Law of Diminishing Returns is not applicable.

Some properties of the aggregate supply function

The AS equation is simply the price determination model Equation (16.2), which shows that in the short run, the behaviour of the aggregate supply in the economy depends on the mark-up ((m), the money wage rate (W), labour productivity (γ)) and raw material costs per unit of output. Here we focus on changes in m, W and γ. Accordingly:

- If the money wage rate rises, other things being equal, the unit cost of production rises and firms would translate this in time into a price rise, thereby restoring the previous mark-up.
- If there is growth in labour productivity (γ), say because of increased labour force morale, increased skill levels, more technologically based production techniques, better management, and the like, then unit costs (W/γ) will fall. This means that the firms can generate the same profit margin at lower prices. The AS function would thus shift downwards by the extent of the decline in marked-up unit costs.
- Variations in the mark-up (m) will cause the price level to change. Increases in industrial concentration, more advertising and such may lead to firms being able to increase the overall profit margin to a level that can be sustained. Tight conditions in the goods and services market, where sales are constrained, may

lead firms to reduce the mark-up as they all struggle for market share. This could also occur if strong trade unions successfully push for money wage increases. Thus, to avoid losing market share, the firms may choose to absorb some of the cost rises by not raising prices, which means that the mark-up has been squeezed.

- If employment is below full employment, then actual output is less than Y^*, which means there is an **output gap**. Increases in aggregate demand (spending) that are seen by firms to be permanent will result in an expansion of output without any price increases occurring. If the firms are unsure of the durability of the demand expansion, they may resist hiring new workers and utilise increased overtime instead. In other words, they initially respond to the increased aggregate spending by increasing hours of work rather than persons employed. The higher costs associated with paying overtime rates are likely to be absorbed in the profit margin because firms want to maintain their overall market share.

The aggregate supply function is a useful vehicle for exploring an inflationary process arising from conflict between groups over the distribution of income. We postpone this analysis until Chapter 17.

16.5 What Determines the Level of Employment?

A firm will hire according to the demand for its services and its demand for labour will not be very sensitive to wage changes. However, it will make decisions about the viability of its operations based in part on wage costs. But on a day-to-day basis, if it is profitable at the current wage rates, then it will increase or decrease its demand for labour based on its expected sales.

In other words, **effective demand drives labour demand**. Firms hire the number of workers they need to produce the amount of output they think they can sell at a profit.

If there has been a prolonged downturn, then we would observe idle capital and labour (unemployment). The unemployed workers are willing to work at the current wage rates but there is no demand for their services because effective demand is too low.

Fixed factor input proportions mean that firms face constant unit costs over the normal range of production, assuming that money wages are fixed in the short run and labour productivity is constant.

If the firm receives increased orders for its output, then it will seek to maintain its market share by increasing output. Assuming constant unit costs, the firm will bring its idle capital back into production and hire more workers. There would be no pressure on the firm to raise prices because there would be no upward pressure on per-unit costs. As output rises, there is increased demand for labour at the constant money wage. This suggests that the AS curve expressed as a function of the price level is very flat over the normal range of output. Increases in nominal demand will be met by increases in output (income).

There are several reasons why firms might be reluctant to increase prices (even though costs might rise temporarily, as we explain below) or reduce them when aggregate demand falls.

- Industries are characterised by a few dominant firms that exercise market power.
- Consumer loyalty to products of other firms means that they will not react to a price fall of similar products.
- A price cut would reduce revenue if it did not induce a sufficient number of consumers to switch brands. Further, competitors might match lower prices in order to retain their consumers.
- There are significant costs involved in adjusting prices. Firms have to produce new price tags and catalogues.

16.6 Factors Affecting Aggregate Output per Hour

What factors determine the impact of change in hours of employment on aggregate output? Over the long term, many influences are at work. These include improvements in technology, changes in the average quality of labour through increased education and improved health, and changes in organisational and management skills, which will lead to a steady increase in the level of output that is produced from a given quantity of inputs.

In seeking to understand short-run employment and output determination, we adopt the view that these influences work slowly over time and so we abstract from them in our short-run analysis.

A strong positive relationship between output per hour and the business cycle is observed in the real world. We call this a **procyclical movement** in output per hour, which means that output per unit of labour input increases as the level of production and employment increases. The procyclical pattern of labour productivity (output per hour) means that per-unit costs will decline as employment rises and the economy moves toward full capacity utilisation. However total production costs will obviously rise.

Figure 16.4 shows real output per person in the US manufacturing sector over a 30-year period. The shaded areas are the recessions defined by the US National Bureau of Economic Research (NBER). The behaviour of labour productivity is clearly procyclical. During a recession, when output is falling, productivity falls. This is in contradistinction to the Law of Diminishing Marginal Productivity. The behaviour of real output per hour worked is similarly procyclical. The US behaves in a similar way to all advanced economies with respect to procyclical movements in labour productivity in the manufacturing sector.

The choice of production technology

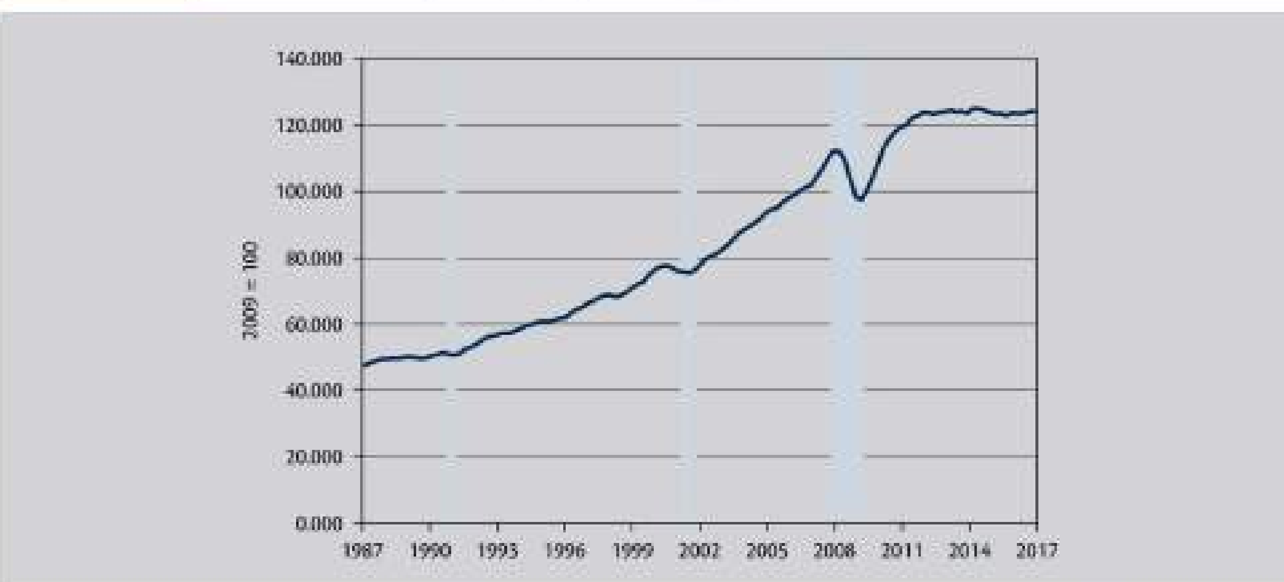
Neoclassical production theory considers that firms can substitute labour and capital freely. Thus, if the relative price of labour increases, firms will quickly use less labour and more capital.

We have seen that firms are rarely able to substitute inputs quickly, and to use more capital and less labour typically requires a total change in technology. The increase in money wages relative to the current price of output would have to be very large to justify the firm scrapping their existing technology.

Consider how firms might act under the fixed input ratio assumption. Based on the available technologies and the projected relative costs of labour and capital into the future, a typical firm will choose the lowest-cost technology. In turn, this will set the capital-labour input ratio by which it will be bound in the coming production periods. In making that decision, the firm is also committing to a certain labour demand given the relationship between the technology being used and the associated input proportions.

Once installed, capital becomes what economists call 'a free good'. Relative to its purchase and installation costs, the variable costs of running the capital are usually low. Economists refer to sunk costs in respect of the

Figure 16.4 US manufacturing output per person employed, 1987 to 2017



Source: Authors' own. Data from U.S. Bureau of Labor Statistics, Manufacturing Sector: Real Output Per Person [PRS30006163], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/PRS30006163, August 6, 2018.

Note: Shaded areas indicate US recessions.

major costs of acquiring capital equipment, which means that the firm has already incurred them whether it runs the plant or not.

Accordingly, the firm will use as much capital as is required to produce the current output that is being demanded. When demand falls, the firms simply leave some proportion of their capital stock idle (underutilised).

But in doing so, they shed labour and/or reduce working hours because the variable costs of the labour input are relatively high when compared to the fixed hiring and related costs.

What role does the real wage play in this? Even if the real wage fell to zero, the firms would not employ more workers if aggregate demand didn't justify it. Firms will not produce if there is not a prospect of sale (barring the small proportion of production they keep as inventories to smooth out orders).

Procyclical movements in labour productivity

As noted, firms will leave machines idle in the light of low demand for their product or services. When the demand for their products and services rises, firms will first try to utilise existing staff and capital more fully. Thus workers will be encouraged to work faster and/or longer, so that productivity per worker rises. So, in the short run, costs might rise as overtime premiums are paid while firms decide whether the increase in demand is permanent or transitory. If the rise in demand is sustained, firms will then increase staff, and if necessary, invest in additional capital equipment. Unit labour costs will fall once new staff are hired and overtime declines. On the other hand, in downturns, firms do not fully adjust the workforce down because they do not wish to lose experienced workers.

Consequently, employment tends to fluctuate less than output or production, with labour being hoarded to some degree in a recession and utilised more intensively at the beginning of a boom. Thus labour productivity is procyclical, which accords with the evidence depicted in Figure 16.4.

Further, if we return to our example of the cleaning firm, as it obtains more contracts for servicing it may first require the existing workforce to clean more offices per day, perhaps by a 'speed up'. We would observe a rise in labour productivity. At some point, the firm must hire an additional worker (and add a broom). Measured labour productivity might be lower for a while until the firm adds enough contracts to fully utilise the larger workforce.

For these reasons, we observe a cyclical component to labour productivity that is inconsistent with diminishing marginal productivity under neoclassical production theory.

Also, it is implausible that labour productivity for the economy as a whole would rise and fall over short periods to the degree that we observe if it were due to supply side factors such as technological change or improved education and training of the workforce. These factors may have an impact on labour productivity in the long run.

We have noted that fluctuations of demand can contribute to the observed procyclical labour productivity in manufacturing. The following factors can also contribute to the observed pattern of labour productivity.

First, the composition of aggregate output is important because the value of output per hour of employment varies considerably among the firms and industries that comprise the total economy. For example, the manufacture of high-tech electrical goods would have a much greater output per hour of labour input than the provision of hairdressing services.

Even if diminishing returns were operating at the individual firm level (by assumption, not a fact), it is most unlikely that countercyclical productivity growth would be observed at the aggregate level.

The composition of output across industries changes as output increases, but the changes are such that the industries where diminishing returns are most apparent experience a reduced share of total output. Then labour productivity can increase as total output rises.

Second, and of great practical importance, is the observation that most firms desire to maintain long-term relations with their labour forces. The reason for this behaviour relates to the fixed costs of hiring (recruiting, screening, training and redundancy provisions) and to the need to maintain morale among the workers.

Efficiency is crucially dependent on the feelings that the workers have towards job security and the like. Firms are also reluctant to dismiss specialised workers for fear of losing them permanently. Likewise, cutting money wages will redistribute total revenue towards profits if prices are maintained, but will damage relationships with incumbent workers who may quit when other job opportunities become available. Widespread wage cuts might damage aggregate demand as workers will have less income to spend.

Over the longer run, if demand continues to grow, firms will invest in plant and equipment, that incorporates the latest technology, which typically increases labour productivity (fewer labour hours to produce the same amount of output), in addition to increasing production capacity. Firms will also engage in other activities that raise labour productivity, such as reorganising the workplace and improving time management. Labour productivity will increase. In addition, rising unit labour costs due to rising wages will provide the incentive for firms to increase labour productivity over longer periods through such initiatives. Thus, rising wages can spur research and development that leads to innovations in technology.

Orthodox economists tend to attribute both the short-run and long-run trends of labour productivity to supply side factors and ignore the major impact of aggregate demand. This is neither theoretically sound, nor does it explain the empirical data on labour productivity over time.

Conclusion

In this chapter, we have developed a mark-up pricing model, which we argue is representative of the price setting behaviour of many firms in a modern monetary economy. This is not to deny that pricing of products, such as fruit and vegetables, is market driven so that shortages or gluts are reflected in current prices.

The chapter has also provided some insights as to how firms behave when market conditions change. A key point is that there are good reasons why firms do not typically attempt to vary the money wages of their employees according to prevailing labour market conditions. This is at odds with the conclusions of the Classical model, which views price and wage flexibility as essential to the achievement of full employment. This view resonates with many economists today, despite the fact that it is based on a fallacy of composition.

Reference

Langbraaten, N., Nordbø, E.W. and Wulfsberg, F. (2008) "Price-setting Behaviour of Norwegian Firms – Results of a Survey," *Norges Bank Economic Bulletin* 79(2), 13–34.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

PART D

UNEMPLOYMENT AND INFLATION: THEORY AND POLICY

CHAPTER

17 UNEMPLOYMENT AND INFLATION

Chapter Outline

Learning Objectives

- Understand how inflation is defined and the sufficient condition for its persistence.
- Understand the nature of inflation as emanating from a conflict over the distribution of national income.
- Understand the basics of the Quantity Theory of Money and its shortcomings.
- Learn why incomes policies have been proposed to control inflationary spirals.

17.1 Introduction

In this chapter, we will review the concept of inflation and discuss various approaches that seek to explain it. An inflationary process can be understood within a general framework whereby different claimants of real GDP and national income struggle to assert their aspirations. In this sense, we cast inflation within the general distributional struggle or conflict that is characteristic of capitalist economies, between workers seeking to maintain or achieve a higher wage and firms seeking to maintain or raise their profit rate.

We will differentiate between **cost push** and **demand pull** as initiating causes of an inflationary process. The first type has been termed cost push inflation because it originates from the costs of production increasing and pushing up the price level. The second type is termed demand pull, because excess nominal demand (relative to output capacity) initially pushes up the price level.

We then consider the Classical Quantity Theory of Money in more detail. This model asserts that there is a direct relationship between money supply growth and inflation, such that the inflationary process is always due to the central bank allowing this growth rate to be excessive. The Quantity Theory is a central element of Monetarism, which we discuss later in the chapter. We show that the basis of the theory is an accounting identity. However, the theory fails in its attempt to demonstrate causality.

In Chapter 18, we will use the ideas presented here to consider the major theoretical and policy debates within macroeconomics with respect to inflation.

17.2 What is Inflation?

There are misconceptions as to what price inflation actually is. An increase in prices is a necessary but not sufficient condition for an inflationary process to unfold. Thus, a negotiated pay increase for workers, or firms increasing their prices to try to increase profit, or a rise in local prices of imported goods following a depreciation of the exchange rate, may or may not initiate an inflationary process.

Inflation is the continuous rise in the price level, so the price level has to be rising for a number of time periods. A one-off price rise is not an inflationary episode.

If the price level rises by ten per cent every month for example, then we would be observing an inflationary episode. In this case, the inflation rate would be considered stable with the price level rising at a constant rate per period.

If the price level was rising by 10 per cent in month one, then 11 per cent in month two, then 12 per cent in month three and so on, then we would be observing an accelerating inflation rate. Extreme cases of **accelerating** inflation are referred to as **hyperinflation**. There have been few instances of this problem in recorded history, but the Weimar Republic in 1920s' Germany and Zimbabwe at the beginning of the 21st century are notable examples. They were marked by a dramatic contraction of the supply potential of the respective economies prior to the hyperinflation (see Chapter 21 for more on this).

Alternatively, if the price level was rising by ten per cent in month one, nine per cent in month two and so on, then the rate of inflation is falling or **decelerating**. If the price *level* starts to fall, then the growth of the price level is negative and this would be a **deflationary** episode.

REMINDER BOX

You may wish to refresh your understanding of the measurement of the consumer price index (CPI) and the computation of the inflation rate by referring back to Chapter 4, Section 4.8.

We can define a normal price level as being the prices that firms are willing to charge when they are operating at normal capacity and earning a profit rate that satisfies their strategic aspirations. (See the discussion of mark-up pricing in Chapter 16.) However, the economic cycle fluctuates around these normal rates of capacity utilisation and firms not only adjust to the flux and uncertainty of aggregate demand by adjusting output, but in some cases, will vary prices. This is particularly the case during a recession.

When there are very depressed levels of activity, firms might offer discounts in order to increase sales and hence capacity utilisation. Thus, they temporarily suppress their profit margins in order to try to raise their respective market shares when overall demand is falling. As demand conditions become more favourable, firms start withdrawing the discounts and prices return to those levels that offer the desired rate of return at normal rates of capacity utilisation. We do not consider these cyclical adjustments in prices to constitute inflation.

17.3 Inflation as a Conflictual Process

Conflict theory situates the problem of inflation as being intrinsic to the power relations between workers and capital (class conflict), which are mediated by government within a capitalist system. It brings together social, political, and economic considerations in a generalised view of the inflation cycle. This mediation by government varies over the course of history but in more recent times has been biased towards protecting the interests of capital, particularly financial capital, at the expense of workers' real wage aspirations.

Conflict theory is most closely identified with inflationary processes initiated by cost push. However, it is important to recognise that an inflationary process, whether initiated by the forces of cost push or demand pull, by definition requires 'two to tango', so that an increase in prices is ongoing. Otherwise the change in the level of wages or prices is a one-off event. The nature of the power relations between workers and capital is integral to understanding all inflationary processes.

In product markets, firms have price setting power and set prices by applying a mark-up to costs. Firms seek to achieve target profit rates that satisfy their shareholders or owners, and these are expressed by the size of the mark-up on their unit costs. Unit costs are driven largely by wage costs, productivity movements and raw material prices. Shifts in any of these determinants can generate cost increases, which price setting firms may pass on by raising prices.

On the other hand, the bargaining strength of workers will depend on their capacity to mobilise effectively, which is typically through trade union action. The shift to non-standard employment, which can include zero-hours contracts in some countries, including the UK, along with reduced rates of unionisation in many developed economies has reduced the bargaining power of the union movement. In many instances this has been reinforced by anti-union legislation.

When employers are dealing with workers individually, they have more power than when they are dealing with a single bargaining unit (trade union), which represents all workers in their workplace.

Thus, firms and trade unions have some degree of market power (that is, they can influence prices and wage outcomes). They are both assumed to target an **income share** and use their capacity to influence **nominal prices and wages** in order to extract that target share.

In each period, the economy produces a given output (real GDP) which is shared between the groups with distributional claims in the form of wages, profits, rents, interest, taxes and so on. In the initial discussion below, we assume away the other income claimants and concentrate on the split between wages and profits. Later, we will introduce a change in an exogenous claim in the form of a rise in the price of raw materials.

If the desired output shares of the workers and firms are consistent with the available output produced, then there is no incompatibility and there will be no inflationary pressures. The available output would be distributed each period at the prevailing levels of nominal wages and profits which satisfy the respective claimants. However, if the distributional claims are incompatible, then the aggrieved group(s) would seek redress by seeking wage increases (labour) and/or impose price increases (firms). We continue this analysis in the next section.

Cost push inflation

There has been a long line of authors, including **Michał Kalecki**, a Marxist, who have identified inflation as emerging as a result of **distributional struggle** over the available real income.

The dynamic that drives a cost push inflation is seen to arise from the underlying social relations in the economy. This theory of inflation recognises that the two sides of the labour market are likely to have conflicting aims and seek to fulfil those aims by imposing costs on the other party.

The capacity of workers to realise nominal wage gains is considered to be pro-cyclical. That is, when the economy is operating at **high pressure** (high levels of capacity utilisation) workers are more able to secure money wage gains. This is especially the case if they are organised into coherent trade unions, which function as a countervailing force to the power of employers.

For example, reductions in Marx's **reserve army of unemployed** as the economy approaches full employment give workers more bargaining power. Trade unions are more likely to demand higher money wages. Firms may fear prolonged strikes, which will damage them at a time when profits are high.

To protect their market share they are more likely under these circumstances to concede to the workers' demands, knowing that they can in turn use their price setting power to defend their profits by increasing prices (that is, restore the previous mark-up). This can be described as a **battle of the mark-ups**. At that point there is no inflation, just one-off increases in money wages and prices and no change to the distribution of national income between wages and profits.

An inflationary process is instigated and perpetuated if the sum of the distributional claims (expressed in nominal terms, money wage demands and mark-ups) remain greater than the available output measured at current prices and neither bargaining party is prepared to concede to the other by ceasing to pursue higher nominal income. Either a **wage-price or price-wage** mechanism can instigate the inflationary process.

Here the concepts of real wage and/or real profit margin resistance become relevant. A wage-price spiral begins with workers pushing for higher real wages, whereas a price-wage spiral refers to a dynamic where firms initiate the bargaining war by trying to push up their real profit margin.

In a high pressure economy, firms may also initiate an inflationary process by trying to increase their profit margins. Workers may attempt to maintain their previous real wage and will thus respond to the higher price level by seeking an increase in nominal wages. If their bargaining power is strong (which from the firm's perspective is usually measured in terms of how much damage the workers can inflict on output and hence profits via industrial action) then they are likely to be successful. If not, they may have to accept the real wage cut imposed on them by the higher price level, which implies that their nominal wages have a reduced capacity to purchase goods and services.

However, if firms are not willing to absorb the squeeze on their profits following the money wage increase, then they will raise prices again and the beginnings of a wage-price spiral occur. If this process continues, then cost push inflation is the result.

The wage-price spiral could develop into a **wage-wage-price spiral** if one section of the workforce seeks to restore relativities after another group of workers succeeds in their nominal wage demands.

The role of government is also implicated. While it is the distributional conflict which initiates the inflationary spiral, government policy has to be compliant for the nascent inflation to persist.

Business firms will typically access credit (for example, overdrafts) to **finance** their working capital needs in advance of realisation of revenue via sales. In an inflationary spiral, as workers seek higher nominal wages, firms will judge whether the costs of industrial action in the form of lost output and sales are higher than the costs of accessing credit to fund the higher wages bill. Typically, the latter option will be cheaper.

If credit conditions become tighter and thus loans become more expensive, then firms will be less able to pay the higher money wages demanded by workers. The impact of the higher interest rates may thus lead to a squeeze on real wages with the consequent negative impact on consumption spending. Firms will also be less willing to invest in new projects given that the cost of funds is higher.

As a consequence, if monetary policy becomes tighter, there will be some point where output growth declines and the workers who are in weaker bargaining positions are laid off. The rising unemployment in turn eventually discourages the workers from continuing to pursue their demand for wage increases and in time the inflationary process would be choked off.

Cost push theory thus hypothesises a trade-off between inflation and unemployment.

The alternative policy stance is for the central bank to accommodate the inflationary struggle by leaving its monetary policy settings (interest rates) unchanged. This accommodation would also likely see the fiscal authorities maintaining existing tax rates and spending growth.

The commercial banks would continue to extend loans and in the process create deposits in the accounts of its business clients. The central bank would then ensure that there were sufficient reserves in the banking system to maintain stability in the payments system. The nominal wage-price spiral would thus fuel the demand for more loans with little constraint.

There are also strong alignments between the cost push theory of inflation and Hyman Minsky's financial instability notion (see Chapter 26). Both theories consider that the behavioural dynamics change across the economic cycle. When economic activity is strong, the banks are more willing to extend credit to those who previously had been considered to be marginal borrowers, and are now seen to be more creditworthy because economic conditions have improved. Equally, firms will be more willing to pass on nominal wage demands because it becomes harder to find labour, and the costs of an industrial dispute in terms of lost sales and profits are high. Workers also have more bargaining power due to the buoyant conditions.

At low levels of economic activity, falling sales and rising unemployment militate against both profit push and wage demands. Also loan delinquency rates tend to be higher and banks become more conservative in their lending practices.

Another example of cost push pressure might come from an increase in the price of a significant imported raw material, such as oil. We will examine this dynamic in the next section.

Keynes also suggested that inflation could arise due to **cost push factors** (also called sellers' inflation). Within the Keynesian tradition, Abba Lerner's *Economics of Employment* (1951) has a coherent discussion of how distributional struggle may lead to a wage-price spiral and generalised inflation as each party seeks to defend their income.

Lerner showed that the dynamic for this wage-price spiral could also result from capital seeking to expand its share of income by pushing up the mark-up on unit costs. Such a strategy could only be successful if workers conceded the real wage cut implied by the higher prices. Firms would be more likely to attempt this strategy when they perceived the bargaining power of workers to be weak, that is, when the unemployment rate was higher. In this way, Lerner recognised that high inflation and high unemployment could co-exist, and thus identified the phenomenon that subsequently became known as **stagflation**.

Raw material price increases

Until now we have been concentrating on workers pursuing nominal wage increases in order to gain higher real wages and/or firms pushing profit margins up to gain a greater profit share of income as the main drivers of an inflationary process.

However, **raw material price shocks** can also trigger cost push inflation. These cost shocks may be imported (for example, an oil-dependent nation might face higher energy prices if world oil prices rise) or domestically sourced (for example, a nation may experience a drought which increases the costs of food crops and impacts on all food processing industries).

TRY IT YOURSELF

Let us consider the example of a situation where there is a price rise for an essential imported resource. The imported resource price shock amounts to a loss of real income for the nation in question. Thus, there is less real income to distribute to domestic claimants.

The question then is who will bear this loss? With less real income being available for distribution domestically, the reactions of the claimants are crucial to the way in which the economy responds to the higher cost of the imports. The loss has to be shared or borne by one of the claimants or the other. What do you think are the strategies available to the various contestant claimants? Which do you think are most likely to be effective?

If, in response to the fall in their profit margins (mark-ups), domestic firms pass on the raw material cost increases in the form of higher prices, then workers would endure a cut in their real wages.

If workers resist this erosion of their real wages and push for higher nominal wage growth, then firms can either accept the squeeze on their profit margins or resist.

The government can employ a number of strategies when faced with this dynamic. It can maintain the existing nominal demand growth, which would be very likely to reinforce the spiral.

Alternatively, it can use a combination of strategies to discipline the inflation process including the tightening of fiscal and monetary policy to create unemployment (the NAIRU strategy), the development of consensual incomes policies and/or the imposition of wage price guidelines (without consensus) (see below).

Ultimately, if the claimants of real income continue to try to pass on the raw material price rise to each other, then it is likely that contractionary government policy will be introduced and unemployment will rise.

A better strategy would be to either change production processes in order to reduce the use of the expensive imported resource, or to find a domestic alternative.

Conflict theory of inflation and inflationary biases

A series of articles in the journal *Marxism Today* in 1974 illustrated the proposition that inflation was the result of a distributional conflict between workers and capital. These articles were written with reference to the early 1970s, when inflation rates rose in many Western economies.

One article by Pat Devine stated that the inflation process was a structural construct embedded in the intrinsic capital labour conflict. He argued that the increased bargaining power of workers (that accompanied the long period of full employment in the post-Second World War period) and the declining productivity growth in the early 1970s imparted a structural bias towards inflation which was manifested in the inflation breakout in the mid-1970s that "ended the golden age."

He further claimed that the prolonged growth of money wages was "unprecedented in the history of capitalism" (Devine, 1974: 80). Capitalists increased prices to maintain profitability and thus countered the attempt to raise real wages.

Large, oligopolistic firms with price setting power engaged in non-price competition (for example, product quality). These firms, however, were interdependent because their market shares were sensitive to their pricing strategies. When a firm was faced with nominal wage demands, its management knew that its rivals would face similar pressure and that their competitive positions would not depend on the absolute price level while the government continued to ensure that effective demand was sufficient to maintain full employment. On the other hand, a firm could lose market share if it increased prices while other firms maintained lower prices. As a result, firms had little incentive to resist the wage demands of their workers and strong incentives to protect their profits by passing on the demands in the form of higher prices.

This structural depiction of inflation as being embedded in the class dynamics of capital and labour, both of which had increased capacity to set prices and defend their real shares of income, implicates Keynesian-style approaches to full employment.

There was also an international component to the structural theory. It was argued that the Bretton Woods system (see Chapter 9) imparted deflationary forces on economies that were experiencing strong domestic demand growth. As national income rose and imports increased, central banks were obliged to tighten monetary policy to maintain the agreed exchange rate parity and the constraints on monetary growth acted to choke off incompatible claims on the available income.

However, when the Bretton Woods system of convertible currencies and fixed exchange rates collapsed in 1971, the structural biases towards inflation came to the fore with floating exchange rates.

Devine (1974: 86) argued that:

> *floating exchange rates have been used as an additional weapon available to the state. Given domestic inflation, floating rates provide a degree of flexibility in dealing with the resultant pressure on the external payments position. However, if a float is to be effective in stabilising a payments imbalance it is likely to involve lower real incomes at home. If a reduction in real wages (or their rate of growth) is not acquiesced in there will then be additional pressure for higher money wages and if this cannot be contained the rate of inflation will increase and there will be further depreciation.*

The structuralist view also noted that the mid-1970s crisis, which marked the end of the Keynesian period, was not only marked by rising inflation but also by an ongoing profit squeeze due to declining productivity growth and increasing external competition for market share. The profit squeeze led to firms reducing their rate of investment (which reduced aggregate demand growth), which combined with harsh contractions in monetary and fiscal policy, created the stagflation that bedevilled the world in the second half of the 1970s.

The resolution to the **structural bias** proposed by economists depended on their ideological persuasion. On the one hand, those who identified themselves as Keynesians proposed incomes policies (which we shall explore in more detail later in this chapter) as a way of mediating the distributional struggle and achieving nominal income claims that were compatible with the available output.

On the other hand, the emerging Monetarists considered the problem to be an abuse of market power by the trade unions and this motivated demands for policymakers to legislate to reduce the bargaining power of workers. The rising unemployment was also not opposed by capital because it was seen as a vehicle for undermining the capacity of the trade unions to make wage demands.

From the mid-1970s, the combined weight of persistently high unemployment and increased policy attacks on trade unions in many advanced nations reduced the inflation spiral as workers were unable to pursue real wages growth, and productivity growth outstripped real wages growth. As a result, there was a substantial redistribution of income towards profits during this period.

The rise of Thatcherism in the UK and Reaganomics in the USA exemplified the increasing dominance of the Monetarist view in the 1980s.

Demand pull inflation

While economists distinguish between cost push and demand pull inflation, the demarcation between the two types of inflation is not as clear cut as one might think.

Demand pull inflation refers to the situation where prices start accelerating continuously because nominal aggregate demand growth outstrips the capacity of the economy to respond by expanding real output.

We have learned from the national accounts that aggregate demand is always equal to GDP, which is the market value of final goods and services produced in some period. We represent that as the product of total real output (Y) and the general price level (P), that is, PY. It is clear that if there is growth in nominal spending (that is, GDP) that cannot be met by an increase in output (Y) then the general price level (P) has to rise.

The dominant view of inflation in the 1960s was based on Keynes' notion of an inflation gap, which he outlined in his 1940 pamphlet, *How to Pay for the War: A Radical Plan for the Chancellor of the Exchequer*.

In the *General Theory* (1936), Keynes had developed the notion of effective demand to help understand how an equilibrium corresponding to less than full employment could arise in a monetary economy. He now wanted to show how there would be a transition to a fully employed economy during wartime.

With the onset of the Second World War, large-scale spending programmes were implemented as part of the war effort. Keynes argued that as employment rose, rising household incomes would drive up consumer spending, which would cause inflation to accelerate even if money wage rates were constant.

While Keynes' plan was devised in the context of wartime spending when faced by tight supply constraints (that is, a restricted ability to expand output), the concept of the **inflationary gap** has been generalised to describe situations of excess demand where aggregate demand is growing faster than the aggregate supply capacity can absorb it.

Keynes defined the inflationary gap as an excess of planned expenditure over the available output at pre-inflation or base prices. The pre-inflation benchmark output was that corresponding to the full utilisation of capacity. Thus, if an economy could meet the growth in nominal expected demand by rapidly expanding the capacity to produce goods and services, an inflationary gap would not open.

This idea was distilled into the **demand pull theory of inflation**. Once full employment was reached, then nominal demand growth beyond that level would be inflationary.

Thus inflation would tend to increase when unemployment fell (see Chapter 18 for an analysis of the Phillips curve which posits this type of relationship). The theory claimed that as nominal demand growth pushes the unemployment rate towards its irreducible minimum (frictional unemployment), wage and price inflation would start to rise. In other words, an inflationary gap would be created by the emergence of excess aggregate demand.

There are several factors present in the real world that attenuate these demand effects on the inflation rate. First, firms incur extensive costs when they change prices, which leads to a 'catalogue' (or 'menu') approach whereby firms will forecast their expected costs over some future period and set prices according to their desired return. They then signal those prices in their catalogues and advertising to consumers and stand ready to supply whatever is demanded at that price (up to exhaustion of capacity). In other words, they do not frequently alter their prices to reflect changing demand conditions. Only periodically will firms typically revise their price catalogues.

Second, trust and reliability are important in economic transactions. For example, firms seek to build relationships with their customers that will ensure product loyalty. In this context, firms will not wish to vary prices after they have been communicated to consumers.

Third, firms also resist cutting prices when demand falls because they want to avoid so-called adverse selection problems, whereby they gain a reputation only as a bargain price supplier. Firms value 'repeat sales' and thus want to foster consumer goodwill.

Circumstances change somewhat when the economy approaches full productive capacity. Then the mix between output growth and price rises becomes more likely to be biased toward price rises (depending on the bottlenecks in specific areas of productive activity). At full capacity, GDP can only grow via inflation (that is, nominal values increase only). At this point the inflationary gap is breached.

When the US government prosecuted the Vietnam War effort in the 1960s, the inflation rate began to rise. In the late 1960s and early 1970s, the demand pull pressures of the spending associated with the war effort combined with sharp rises in oil prices following the formation of the Organisation of Petroleum Exporting Countries cartel (OPEC). OPEC's oil prices quadrupled in 1973 and generated huge cost shocks to oil-dependent economies such as the US and Japan.

Cost push and demand pull inflation: a summary

Cost push inflation requires certain aggregate demand conditions for it to be sustained. In this regard, it is hard to differentiate between an inflationary process which was initiated from supply side pressures from one that was initiated by demand side pressures.

For example, an imported raw material shock means that a nation's real income that is available for distribution to domestic claimants is lower. This will not be inflationary unless it triggers an ongoing distributional conflict as domestic claimants (workers and capital) try to pass on the real loss to each other.

However, that conflict needs 'oxygen' in the form of ongoing economic activity in sectors where the spiral is robust. In that sense, the conditions that will lead to an accelerating inflation – high levels of economic activity – will also sustain an inflationary spiral emanating from the demand side.

17.4 The Quantity Theory of Money

As we saw in Chapter 11, the Classical theory of employment is based on the view that the real variables in the economy – output, productivity, real wages, and employment – are determined by the equilibrium outcome in the labour market.

> **By way of summary, the real wage is determined exclusively by labour demand and labour supply, which also determine the real level of economic activity at any point in time.**

Say's Law, which follows from the loanable funds doctrine (see Chapter 11), is then invoked to assume away any problems in matching aggregate demand with this supply of goods and services. Under this doctrine saving and investment will always be brought into balance by movements in the interest rate, which is construed as being the price of today's consumption relative to future consumption. Thus two relative prices – the real wage in the labour market and the real interest rate in the loans market – ensure that full employment occurs (with zero involuntary unemployment).

This separation between the explanation for the determination of the real economic outcomes and the theory of the general price level is referred to as the **classical dichotomy**, for obvious reasons. The later Classical economists believed that if the supply of money is doubled, for example, there would be no impact on the real performance of the economy. All that would happen is that the price level would double.

The classical dichotomy that emerged in the 19th century stands in contradistinction to the earlier ideas developed by economists such as David Hume that there is a trade-off between unemployment and inflation that could be manipulated (in policy terms) by the central bank varying the money supply (Hume, 1752).

It is of no surprise that the Classical employment model relies in part on the notion of a dichotomy for its conclusions. Its origins were based on a barter model in which there is an absence of money and owner-producers trade real products. Clearly, this conception of an economy has no application to the monetary economy we live in.

Classical monetary theory was only intended to explain the level and change in the general price level. The main attention of the Classical economists was in trying to understand the supply of output and the accumulation of productive capital (and hence economic growth).

The theory of the general price level that emerged from the Classical dichotomy was called the **Quantity Theory of Money**, which was outlined in Chapter 11. The theory had its origins in the work of French economists in the 16th century, in particular, Jean Bodin.

Why would we be interested in something a French economist conceived in the 16th century? The answer is that just as the main ideas of Classical employment theory still resonate in the public debate (for example, the denial that mass unemployment is the result of a deficiency of aggregate demand), the theory of inflation that arises from the Quantity Theory of Money is still influential. Indeed, it forms the core of what became known as **Monetarism** in the 1970s.

As we have learned already from this textbook, economics is a contested discipline and different schools of thought advance conflicting policy frameworks. Monetarism and its more modern expressions form one such school of thought in macroeconomics and rely on the Quantity Theory of Money for their inflation theory.

We will also see that the crude theory of inflation that emerges from the Quantity Theory of Money has intuitive appeal and is not very different to what we might expect the average layperson to believe: that growth in the money supply causes the value of money to decline (that is, causes inflation).

The Quantity Theory of Money was very influential in the 19th century. The theory begins with what was known as the **equation of exchange**, which is an accounting identity. We write the equation as:

$$(17.1) \qquad M_s V = PY$$

You are familiar with the terms on the right-hand side. PY is the nominal value of total output (which is simply the definition of nominal GDP in the national accounts) given that P is the price level and Y is real output.

M_s is the quantity of money in circulation (the money supply, say M2 which was defined in Chapter 10), which is a stock (so many dollars at a point in time). V is called the income velocity of circulation, and is the average number of times the stock of money turns over in the generation of aggregate income.

There is no theoretical content in the Equation (17.1) as it stands, since it is an identity. We thus need to introduce some behavioural elements in order to use Equation (17.1) as a theory of the general price level.

BOX 17.1 VELOCITY EXAMPLE

To understand velocity, we can consider the following example of an imaginary and simple economy. Assume the total stock of money is \$100, which is held by the two people that make up this economy. In the current period (say a year), Person A buys goods and services from Person B for \$100. In turn, Person B buys goods and services from Person A for \$100.

The total transactions equal \$200 yet there is only \$100 (money stock) in the economy. Thus each dollar must be used twice over the course of the year. So the velocity in this economy is two.

The velocity of circulation converts the stock of money into a flow of monetary spending and renders the left-hand side of Equation (17.1) commensurate with the right-hand side.

In this regard, it is important to see the Quantity Theory of Money and Say's Law as being mutually reinforcing planks of the Classical theory. Say's Law was proposed to justify the presumption that full employment output would be continuously supplied and sold, which meant that the Quantity Theory of Money would ensure that changes in the stock of money would only impact on the price level.

As Keynes observed, price level changes do not necessarily correlate with changes in the money supply, and this led to his rejection of the Quantity Theory of Money. Another way of stating this is that the velocity of money need not be fixed, and real output need not tend to the full employment level.

In turn, Keynes' understanding of how the price level could change without a change in the money supply was informed by his rejection of Say's Law. He recognised that total employment is determined by effective demand and that a capitalist monetary economy could experience deficient effective demand.

However, the Classical theorists considered that a flexible real wage would ensure that full employment is attained, at least as a normal state where competition prevails and there are no artificial real wage rigidities imposed. As a result, they considered *Y* to be fixed at the **full employment output level**.

Additionally, they considered *V* to be constant given that it is determined by customs and payment habits. For example, people are paid on a weekly or a fortnightly basis and shop say, once a week for their needs.

Equation (17.2) depicts the resulting causality that defines the Quantity Theory of Money as an explanation of the general price level. The horizontal bars above the *V* and *Y* indicate that they are assumed to be constant. It follows that changes in M_s will directly and only impact on *P*.

$$(17.2) \qquad M_s\overline{V} = P\overline{Y}$$

$$\therefore M_s \rightarrow P$$

To understand this theory more deeply it is important to note that the Classical economists considered the role of money to be confined to acting as a medium of exchange to free people from the tyranny of the necessity of a double coincidence of wants under the barter system. In other words, money would overcome the problem of a farmer who had carrots to offer but wanted some plumbing done, and could not find a plumber desiring any carrots, for example.

Money is thus seen as the means of lubricating the exchange of goods and services. There is no other reason why a person would wish to hold it under this limited conception of money.

The underlying view is that if individuals found they had more money than in the past, then they would try to spend it. Logically, it follows that they consider a rising stock of money to be associated with the growth in aggregate demand (spending).

As Equation (17.2) shows, monetary growth (and the assumed extra spending) would directly lead to price rises because the economy is already assumed to be producing at its maximum productive capacity and the habits underpinning velocity are stable.

For now you should note two empirical facts. First, capitalist economies are rarely at full employment. Since economies typically operate with spare productive capacity and often with high rates of unemployment, it is hard to maintain the view that there is no scope for firms to expand real output when there is an increase in nominal aggregate demand.

Thus, if there is an increase in availability of credit and borrowers use the deposits that are created by the loans to purchase goods and services, firms with excess capacity are likely to respond by raising real output to maintain market share rather than raising prices.

Second, the empirical behaviour of the velocity of circulation demonstrates that the assumption that it is constant is implausible. Figure 17.1 uses data provided by the US Federal Reserve Bank of St. Louis and shows the velocity of circulation, which is constructed as the ratio of nominal GDP to the M2 measure of the money supply.

The US Federal Reserve Bank of St. Louis defines this measure "as the rate of turnover in the money supply–that is, the number of times one dollar is used to purchase final goods and services included in GDP" (2016).

The evidence does not support the claims of the Quantity Theory of Money. No simple proportionate relationship exists between rises in the money supply and rises in the general price level.

Figure 17.1 Velocity of M2 money stock, US, 1950–2015

Source: Authors' own. Data from US Treasury via US Federal Reserve Bank of St Louis. Shaded areas indicate US recessions.

17.5 Incomes Policies

Governments facing a wage-price spiral have from time to time considered the use of so-called **incomes policies** if they were reluctant to introduce a sharp contraction in the economy, which might otherwise discipline the combatants in the distributional struggle.

Incomes policies in general, are measures that are aimed to control the rate at which wages and prices rise, as the economy moves toward, or is at full employment. Progressive economists often advocate their use to rein in cost pressures and avoid the need to reduce overall spending, which creates higher involuntary unemployment.

Incomes policies have been introduced in various forms at various times in a number of countries as a way of reducing supply side cost pressures and allowing employment to stay at a higher level. For example, in 1962, the US government introduced wage price guideposts, which allowed for an average rate of nominal wage increase equal to the average annual rate of productivity growth in the overall economy. This means that per-unit labour costs of production remained constant. Other nominal incomes, including profits, were also to be tied to this rule.

Taken together, it was considered that this rule would stabilise the growth in nominal incomes (and directly link real income growth to productivity growth), thereby reducing any inflationary pressures associated with the maintenance of full employment. Its application would thus distribute productivity gains across all income earners and thus reduce the distributional conflict, which might otherwise instigate a wage-price spiral. However, a problem with the rule is that workers in above-average productivity growth sectors are undercompensated, and workers in below-average sectors are overcompensated. Also, workers would be unable to pursue money wage increases in response to profit pushes by firms.

For a time, the guidelines seemed to work. But as US government expenditure grew as a result of the Vietnam War effort and unemployment fell below four per cent, wage increases began to exceed average productivity growth. By 1966, the guidelines provided no discipline on the growth of nominal incomes in the US. It was clear that the US government was unable to compel employers to follow the guideposts in the wage bargaining process.

Despite the failure of the wage price guideposts, the Republican administration under Nixon reintroduced an incomes policy in 1971. Initially, this was in the form of a 90-day freeze on wages and other nominal incomes. Later, compulsory growth guidelines were set for wages and prices growth.

In 1973, the government introduced yet another freeze on prices, followed by sector-by-sector price rises in line with cost increases with a freeze on profit margins, so workers were exposed to rising prices of oil and food. The experiment ended in April 1974. It was considered a success when it was in place, but when the controls were eliminated, prices and wages began to rise again, although wage and price pressures coming from the demand side were subdued.

The problem was ongoing pressure from the cost (supply) side, in particular from energy and food (largely grains) prices, which led to higher price inflation. Workers were unable to secure money wage increases in line with price inflation, which contributed to the divergence between real wage growth and productivity growth.

On the other hand, in the UK and Australia, the institutional structures that made economies more susceptible to distributional conflict in the late 1960s and early 1970s also made the operation of incomes policies difficult. Highly-concentrated industries, with large firms exercising significant price setting power, were interacting with strong trade unions. These firms were in a strong position to pass on wage demands in the form of higher prices, and governments were reluctant, or unable constitutionally, to mandate strict wage price controls in normal times.

However, incomes policies have worked more effectively in some European nations, for example, Austria and the Scandinavian countries. These nations have long records of collective bargaining and are more attuned to tripartite negotiations than the English-speaking nations. A good example of a successful incomes policy approach, where wages and prices growth were driven by productivity growth in certain sectors, is the so-called **Scandinavian Model** (SM) of inflation (see Box 17.2). This approach to wage setting was developed in Sweden and attempted to marry notions of fairness, the effectiveness of centralised wage bargaining and international competitiveness.

By the late 1970s, incomes policies lost favour in most countries as a result of the rising dominance of Monetarism, which eschewed institutional solutions to distributional conflict in favour of market-based approaches involving higher unemployment.

The Monetarist approach in many advanced nations combined the use of persistently high unemployment with policies designed to reduce the bargaining power of workers. This reduced inflationary pressures because workers were unable to pursue real wages growth and as a result productivity growth outstripped real wages growth. This led to a substantial redistribution of income towards profits during this period. The rise of Thatcherism in the UK exemplified the increasing dominance of the Monetarist view in the 1980s.

In Chapter 19, we will introduce the concept of employment and unemployment **buffer stocks** in a macroeconomy and analyse how they can be manipulated by policy to maintain price stability.

BOX 17.2 THE SCANDINAVIAN MODEL (SM) OF INFLATION

This model, which was originally developed for fixed exchange rates, dichotomises the economy into a competitive sector (C sector) and a sheltered sector (S sector). The C sector produces products which are traded on world markets, and its prices follow the general movements in world prices. The C sector serves as the leader in wage settlements. The S sector does not trade its goods externally.

Under fixed exchange rates, the C sector maintains price competitiveness if the growth in money wages in its sector is equal to the rate of change in its labour productivity (assumed to be superior to S sector productivity) plus the growth in prices of foreign goods. Under this condition, price inflation in the C sector is equal to the foreign inflation rate. The wage norm established in the C sector spills over into wages growth throughout the economy.

The S sector inflation rate thus equals the wage norm less its own productivity growth rate. Hence, aggregate price inflation is equal to the world inflation rate plus the difference between the productivity growth rates in the C and S sectors weighted by the S sector share in total output. The domestic inflation rate can be higher than the rate of growth in foreign prices without damaging competitiveness as long as the rate of C sector inflation is less than or equal to the world inflation rate.

In equilibrium, nominal labour costs in the C sector will grow at a rate equal to the norm (the sum of the growth in world prices and the C sector productivity). Where non-wage costs are positive (taxes, social security and other benefits extracted from the employers) and possibly growing, the requirement is that per-unit variable costs grow at the rate of world prices. The long-run tendency is for nominal wages to absorb the room provided. However, in the short run, labour costs can diverge from the permitted growth path. This disequilibrium must emanate from domestic factors.

The main features of the SM can be summarised as follows:

- The domestic currency price of C sector output is exogenously determined by world market prices and the exchange rate.
- The surplus available for distribution between profits and wages in the C sector is thus determined by the world inflation rate, the exchange rate and the productivity performance of industries in the C sector.
- The wage outcome in the C sector flows on to the S sector industries either by design (solidarity) or through competition.
- The price of output in the S sector is determined (usually by a mark-up) by the unit labour costs in that sector. The wage outcome in the C sector and the productivity performance in the S sector determine the change in unit labour costs.

An incomes policy would establish wage guidelines, which would set national wages growth according to trends in world prices (adjusted for exchange rate changes) and productivity in the C sector. This would help to maintain a stable level of profits in the C sector. Whether this was an equilibrium level depends on the distribution of factor shares prevailing at the time the guidelines were first applied.

Clearly, the outcomes could be different from those suggested by the model if a short-run adjustment in factor shares was required. Once a normal share of profits was achieved, the guidelines could be enforced to maintain this distribution.

A major criticism of the SM as a general theory of inflation is that it ignores the demand side. Uncoordinated collective bargaining and/or significant growth in non-wage components of labour costs may push costs above the permitted path. Where domestic pressures create divergences from the equilibrium path of nominal wage and costs, there is some rationale for pursuing a consensus-based incomes policy.

By minimising domestic cost fluctuations faced by the exposed sector, an incomes policy could reduce the possibility of a C sector profit squeeze, help maintain C sector competitiveness, and avoid employment losses. Significant contributions to the general cost level and hence prices, can originate from the actions of government. Payroll taxation and various government charges may in fact be more detrimental to the exposed sector than increased wage demands from the labour market.

Although the SM was originally developed for fixed exchange rates, it can accommodate flexible exchange rates. Exchange rate movements can compensate for world price changes and local price rises. The domestic price level can be completely insulated from the world inflation rate if the exchange rate continuously appreciates (at a rate equal to the sum of the world inflation rate and C sector productivity growth).

Similarly, if local price rises occur, a stable domestic inflation rate can still be maintained if a corresponding decrease in C sector prices occurs. An appreciating exchange rate discounts the foreign price in domestic currency terms.

What about terms of trade changes? Terms of trade changes, which in the SM justify wage rises, also (in practice) stimulate sympathetic exchange rate changes. This combination locks the economy into an uncompetitive bind because of the relative fixity of nominal wages. Unless the exchange rate depreciates far enough to offset both the price fall and the wage rise, profitability in the C sector will be squeezed.

Policy makers (particularly in Sweden) considered it appropriate to ameliorate this problem through an incomes policy. Such a policy could be designed to prevent destabilising wage movements in response to terms of trade improvements. In other words, wage bargaining, which is consistent with the mechanisms defined by the SM, may be detrimental to both the domestic inflation target and the competitiveness of the C sector and may need to be supplemented by a formal incomes policy to restore or retain consistency.

Conclusion

This chapter is designed to provide an introduction to the concept of inflation, to highlight that it arises due to the conflictual nature of the capitalist system and that ongoing inflation requires that the major combatants (firms and workers) continue to pursue increases in their nominal incomes. The initiating conditions for an inflationary process can be conceptualised in terms of cost push and demand pull, but in practice it is hard to distinguish between them when an outbreak of higher inflation occurs.

We reviewed the Quantity Theory of Money which is based on an identity. When behavioural assumptions are introduced, the theory implies that a simple proportionate relationship exists between increases in the money supply and rises in the general price level. However, no such relationship has been found so, even if it were possible to control the money supply, there would not be a systematic impact on inflation.

Incomes policies were examined, in particular the Scandinavian Model (SM) of inflation. It was noted that they have largely gone out of favour and countries have tended to rely on the use of unemployment as a buffer stock, that is, to rely on higher unemployment to address an inflation rate which is considered to be too high, irrespective of the initial drivers of the inflationary process.

References

Devine, P. (1974) "Inflation and Marxist Theory", *Marxism Today*, March, 79–92.

Hume, D. (1752) "Of Money", in D. Hume (ed.), *Political Discourses*, Edinburgh: Fleming.

Keynes, J.M. (1936) *The General Theory of Employment, Interest, and Money*, London: Macmillan, 1957 Reprint.

Keynes, J.M. (1940) *How to Pay for the War: A Radical Plan for the Chancellor of the Exchequer*, London: Macmillan.

Lerner, A. (1951) *Economics of Employment*, New York: McGraw-Hill.

US Federal Reserve Bank of St. Louis (2016) *Money Velocity*. Available at: https://research.stlouisfed.org/fred2/categories/32242, accessed 20 February 2016.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

18 THE PHILLIPS CURVE AND BEYOND

Chapter Outline

Learning Objectives

- Understand the early Phillips curve debate.
- Recognise that the expectations augmented Phillips curve has had a profound influence on macroeconomic policy in developed economies for over 40 years.
- Gain familiarity with the developments in the Phillips curve literature since the 1980s, including hysteresis and the role of rising underemployment in the inflationary process.

18.1 Introduction

In this chapter, we build on the analysis presented in the previous discussion about inflation (Chapter 17) and introduce the Phillips curve. In the immediate post-Second World War Keynesian era, the concept of full employment was defined as the availability of sufficient jobs to match the preferences of the available labour force. Accordingly, at full employment, no worker who wanted to work would be involuntarily unemployed.

This post-war consensus was steadily eroded away over the next 30 or so years. By the early to mid-1970s, mainstream macroeconomics reverted back to the pre-Keynesian notions of voluntary unemployment and effectively abandoned the concept of true full employment.

However, the process of abandoning true full employment began in the 1950s when the discussion turned to the twin evils of unemployment and inflation. This was the era in which the **Phillips curve** literature emerged, based on what was then considered to be a statistically reliable, inverse relationship between unemployment and inflation, the so-called **trade-off between unemployment and inflation**.

However, later Monetarist and New Classical reinterpretations of the trade-off appeared and became the dominant view as the neoliberal 'free market' school of thought took centre stage and further moved economics away from a coherent notion of full employment. Classical (pre-Keynesian) notions of a natural unemployment rate (equated with full employment) were revived, which led to the rejection of demand management policies that aimed to limit unemployment to its frictional component.

In this chapter, we will carefully analyse the Phillips curve and how the idea that there might be a trade-off between the twin evils of unemployment and inflation has changed with the augmentation of the Phillips curve with inflationary expectations and the so-called **natural rate of unemployment**.

Since the early 1970s, ideological dominance in this debate has been assumed by those who eschew the intervention of government and consider that the unfettered operation of the market will generate full employment. The persistence of mass unemployment around the world is testament to the error of their thinking.

Finally, drawing on empirical evidence, we develop a model of inflation that exploits the concept of hysteresis to justify the restoration of a trade-off between the inflation rate and a broader measure of labour underutilisation.

This approach supports the MMT and post-Keynesian view more generally that the government has an important role to play in maintaining low levels of unemployment.

18.2 The Phillips Curve

In Chapter 16 we derived what we defined as the aggregate supply function (Figure 16.3, p. 247), which was reverse L shaped. The horizontal segment was explained by the price mark-up rule and the assumption of constant unit costs over a range of outputs. In other words, firms in aggregate are assumed to supply as much real output (goods and services) as is demanded at the current price level, up to a limit defined by the available capacity.

The aggregate supply function becomes vertical at full employment because beyond that point the economy exhausts its capacity to expand short-run output due to shortages of labour and capital equipment. Firms will then be trying to outbid each other for the already fully employed labour resources and in doing so will drive money wages up. Under normal circumstances, the economy will rarely approach the output level (Y^*), which means that the economy usually faces constant costs.

We acknowledged in Chapter 16, however, that rising costs **might** be encountered at a lower level of aggregate output, given that all firms or sectors are unlikely to hit full capacity simultaneously so that cost pressures would begin to mount before the overall full capacity output was reached. Thus, the reverse L shaped aggregate supply curve represented an analytical simplification.

It was considered plausible that inflationary pressures rise in inverse proportion to the unemployment rate (which was taken to be a proxy for the overall health of the economy). While earlier researchers had studied this relationship, it was not until 1958 that New Zealand economist **Bill Phillips** published a statistical study that showed the relationship between the unemployment rate and the rate of change (growth rate) of money wages for the UK in the period 1861 to 1957.

Phillips believed that since money wage costs represent a high proportion of total costs, movements in money wage rates would drive movements in the general price level. Later economists constructed the relationship as being between the rate of price inflation and the unemployment rate.

In the Phillips curve framework, the level of economic activity is represented by the unemployment rate, so the model linked the level of economic activity to changes in the price level. Therefore, when the unemployment rate rises above some irreducible minimum, economic activity declines and wage pressures decline. Conversely, as the unemployment rate moves towards that irreducible minimum, the economy moves closer to full capacity utilisation and full employment and wage pressures escalate. Figure 18.1 shows a stylised Phillips curve.

Phillips (1958: 283) explained this empirical relationship by reference to the supply and demand for labour:

> *When the demand for labour is high and there are very few unemployed we should expect employers to bid wage rates up quite rapidly ... On the other hand it appears that workers are reluctant to offer their services at less than the prevailing rates when the demand for labour is low and unemployment is high so that wage rates fall only very slowly. The relation between unemployment and the rate of change of wage rates is therefore likely to be highly non-linear.*

He also recognised that in wage bargaining, the direction of change in the economy was a factor that had to be considered, quite apart from the level of economic activity. Thus, at a particular rate of unemployment, employers would bid more vigorously for the services of labour when demand was increasing than when it was falling.

Note that Figure 18.1 assumes that at very high levels of unemployment, the growth in the money wage is negative. In normal economic conditions there tends to be downward money wage rigidity, and so the Phillips curve would always be associated with positive or (close to) zero money wage growth. However, during the GFC a number of Eurozone economies experienced falling money wages.

Soon after Phillips released his work, American economists Paul Samuelson and Robert Solow (1960) defined the Phillips Curve as a policy tool which the government could use to lessen the burden of unemployment. Reference was made to a **menu of choices**, which implied that policymakers could choose a combination of inflation and unemployment which corresponded to a preferred point on their estimated Phillips curve. This introduced the idea of a **policy trade-off** between unemployment and inflation (see Figure 18.2). If the government wanted to sustain lower unemployment rates, then the cost of that policy decision would be higher inflation.

What would be the best policy choice for government? This was a political decision and governments tried to assess what might be the socially acceptable combination of the twin evils. Depending on the ideological preferences of the voters, some nations might choose Point B in Figure 18.2, where the inflation rate is low but

Figure 18.1 The basic Phillips curve

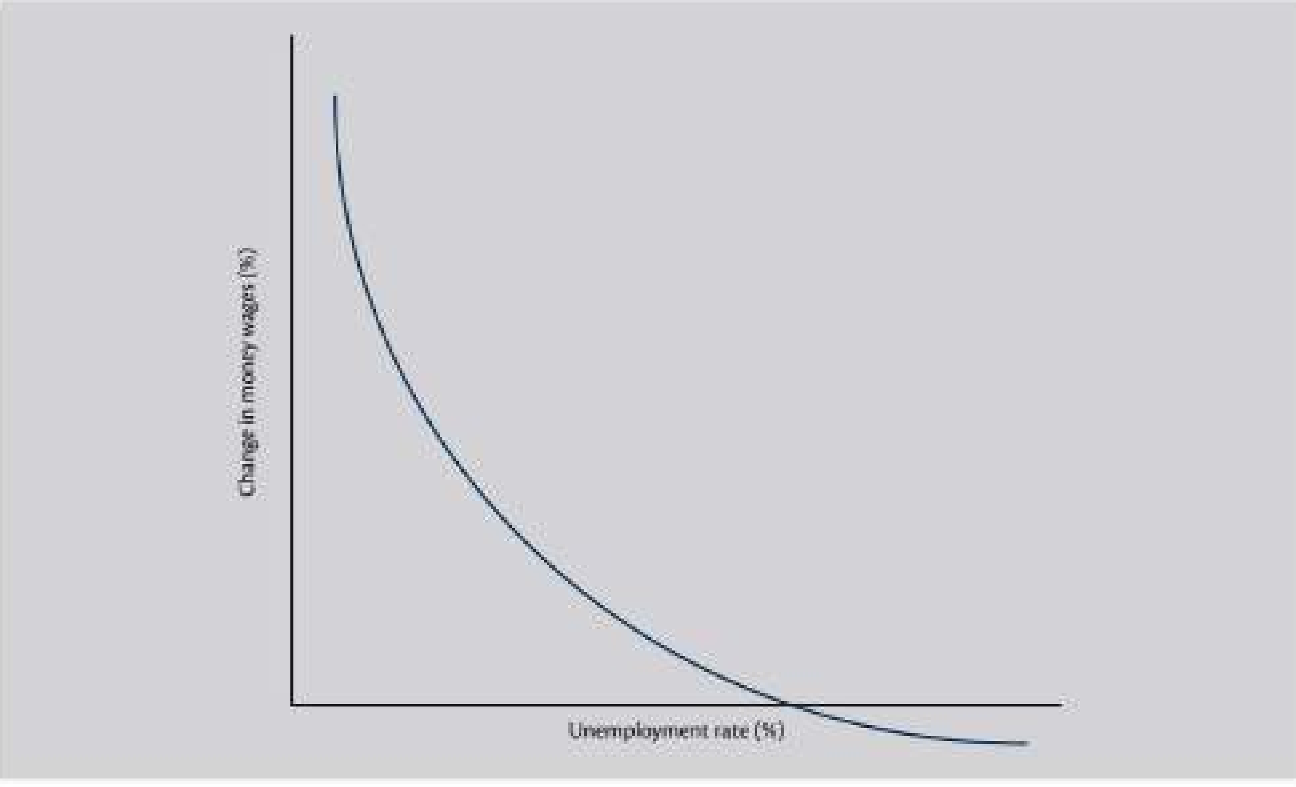
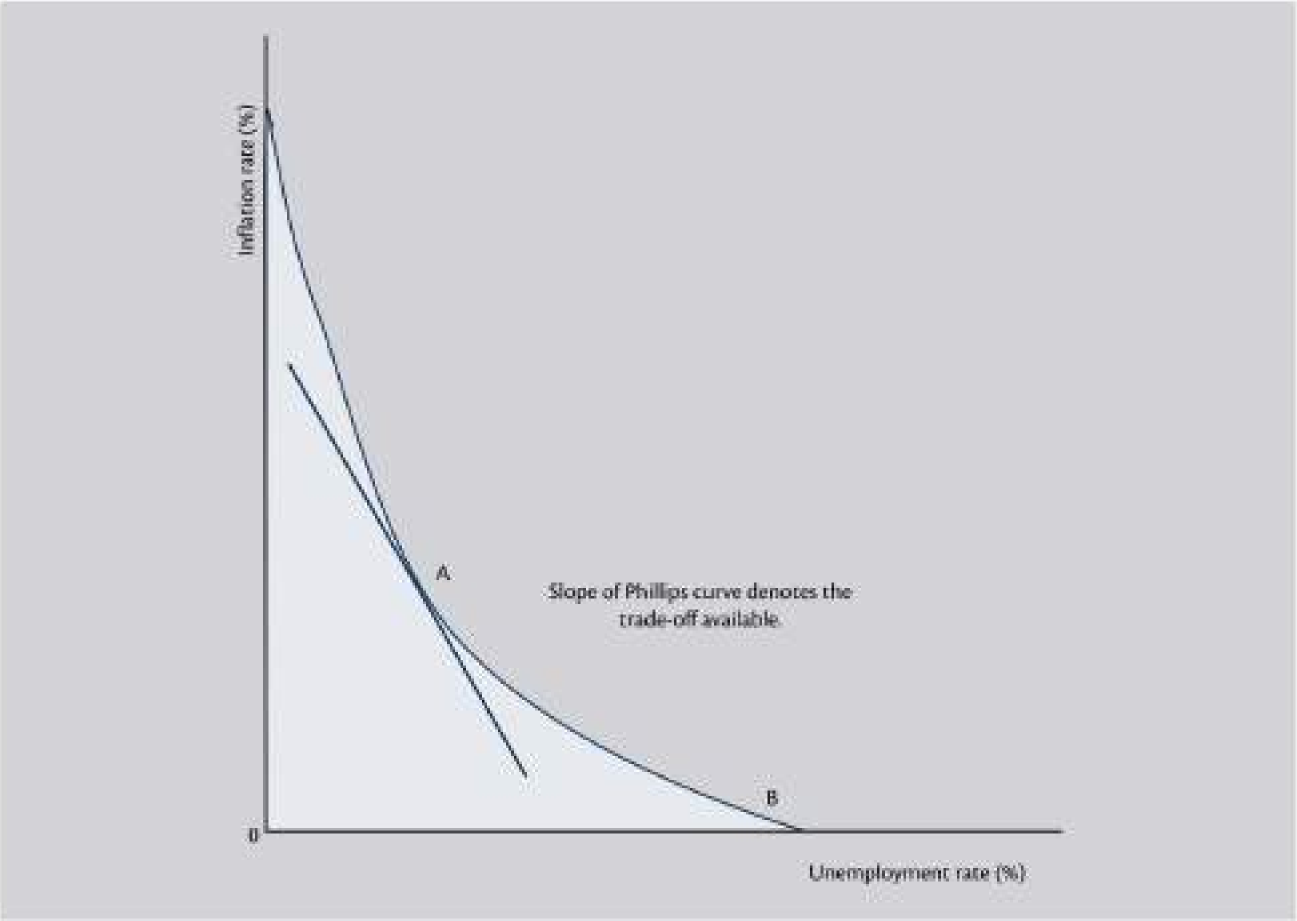


Figure 18.2 The unemployment inflation choice set



unemployment high, while other nations would prefer Point A with higher inflation and lower unemployment. The slope of the Phillips curve at each point indicates the trade-off possibilities.

There would be a political problem for government if its preferred trade-off of inflation and unemployment was unattainable because it lay within the unshaded area beneath the curve rather than on the Phillips curve itself.

Phillips curve algebra

The simplest non-linear form of the original Phillips curve relationship between the growth in money wages and the unemployment rate is given as:

$$(18.1) \quad \dot{W}_t = \alpha_0 + \beta U_t^{-1} \qquad \beta > 0$$

where $\dot{W}_t$ is the rate of money wage inflation (the dot above the W represents a rate of change), $\alpha_0 < 0$ is a constant, β is a coefficient which tells us how the rate of wage inflation responds to the excess demand variable, U_t^{-1}, and the subscript t denotes the current period.

The term βU_t^{-1} measures the impact of the state of the labour market on money wage growth, so that as the rate of unemployment rises, the rate of money wage inflation falls. U_t^{-1} is the inverse of the unemployment rate and captures the non-linear shape of the Phillips curve, as hypothesised by Phillips and later by Samuelson and Solow. The rate of money wage inflation asymptotes to a value of α_0 as the rate of unemployment increases.

We saw in Chapter 15 (*The Aggregate Expenditure Model*) that the convention is to impose positive values on coefficients, such as the marginal propensity to import (m), and to impose appropriate signs on the corresponding terms within the equation, so that imports appear in the aggregate expenditure expression with a negative sign. On the other hand, a constant (intercept) term appears with a positive sign but may be positive or negative. In this non-linear specification the U_t^{-1} term appears with a positive sign and, by convention, parameter β is positive.

We noted in Chapter 17 that when money wages grow in line with labour productivity, there will be no inflationary pressures coming from the labour market. In other words, price inflation (holding other cost factors constant) will be equal to the growth in money wages minus labour productivity growth ($\dot{\gamma}$). Then the price inflation equation can be written as:

$$(18.2) \qquad \dot{P}_t = \dot{W}_t - \dot{\gamma} = \alpha_0 + \beta U_t^{-1} - \dot{\gamma} = \alpha + \beta U_t^{-1}$$

where $\alpha = \alpha_0 - \dot{\gamma} < 0$.

A simpler specification is in linear terms so the money wage Phillips curve is written as:

$$(18.3) \qquad \dot{W}_t = \alpha_0 - \beta U_t \qquad \beta > 0$$

where β is now the sensitivity of the growth of money wages to changes in the unemployment rate.

Then the expression for the price Phillips curve can be written as:

$$(18.4) \qquad \dot{P}_t = \alpha - \beta U_t$$

where again $\alpha = \alpha_0 - \dot{\gamma} < 0$, but $\alpha > 0$.

This also tells us that the rate of general price inflation will be higher, the lower is the unemployment rate and the lower is the productivity growth. If you look at Equations (18.2) and (18.4) you can see that each specification has an excess demand term relating to the rate of unemployment and a constant term.

The instability of the Phillips curve

Consider Figure 18.3, which shows the combinations of the unemployment rate and the annual price inflation rate for the US from 1948 to 2015. The logarithmic regression between the inflation rate and the unemployment rate has been plotted for the three sub-periods: 1948–69, 1970–80, and 1981–2015.

The regression line for the period 1948–69 fits the data quite well, and shows a simple Phillips curve trade-off of the type depicted in Figure 18.2.

However, consider the observations for 1970 to 1980. Those data are clearly inconsistent with a stable Phillips curve and seem to imply a positively sloping relationship. This apparent **shift in the Phillips curve** was considered to be a collapse in the relationship and led to accusations that the underlying conceptualisation of the Keynesian Phillips curve was flawed.

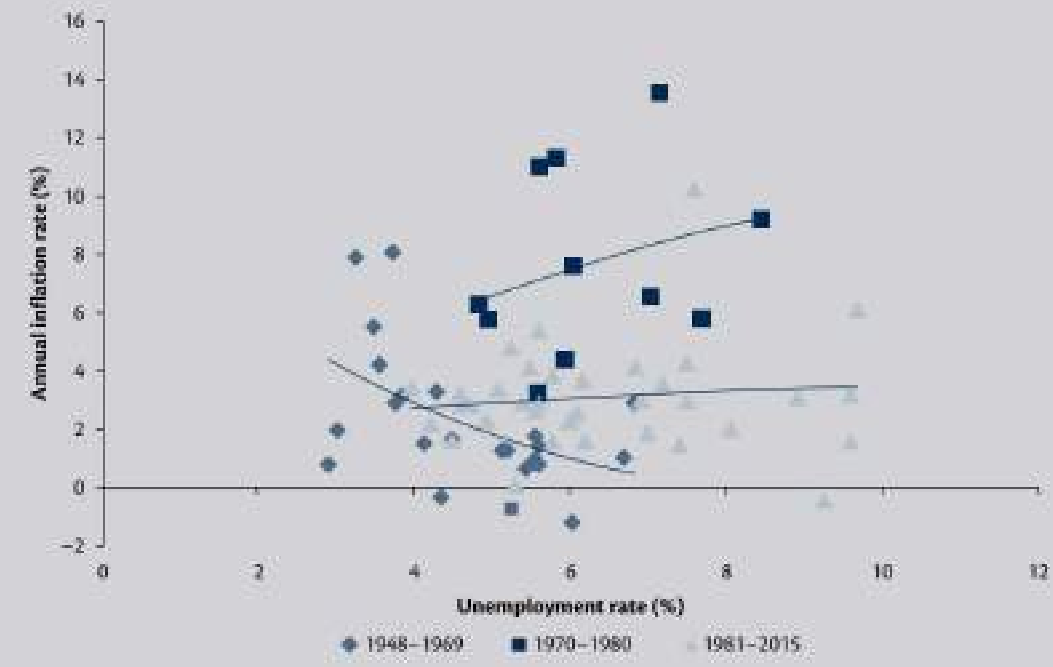
By the 1980s, when inflation moderated, it became hard to determine any relationship between inflation and unemployment in the US economy.

From an empirical perspective then, the belief that the Phillips curve was a stable relationship which could be exploited in a predictable manner by policymakers, according to their preferences between inflation and unemployment, had become highly questionable.

We always have to be very careful when we visualise data in this way. First, the choice of dates for the different periods is important. Second, the observations between 1970 and 1980 may in fact signify a shifting Phillips curve relationship and the regression line is just picking up the shifting function. However, for the trade-off to be empirically sustained, it is necessary for consecutive annual observations to define a trade-off via econometric estimation, and a rationale needs to be provided as to whether or not the trade-off is shifting.

The idea of stagflation that Lerner advanced in the early 1950s (see Chapter 17) would also help to understand the empirical instability in the Phillips curve that began to manifest in the late 1960s and led to a major shift in macroeconomic thinking.

Figure 18.3 The shifting US Phillips curve, 1948–2015



Source: Authors' own. Data from US Bureau of Labor Statistics. The inflation rate is the annual rise in the CPI.

Econometric misspecification

It was known that the Phillips curve became unstable (moved around) in the late 1960s and was particularly susceptible to sudden and/or large increases in inflation. The econometrically estimated consumption functions in the large macroeconomic policy models that were popular in the 1960s also became unstable in the 1970s. Some economists successfully showed that the failure of the large-scale econometric models maintained by governments and central banks in the early 1970s to forecast variables such as savings and consumption could be traced to the misspecification of the consumption function. Most of these models ignored the possibility that rising inflation would influence consumption (for example, if consumers expect prices to rise quickly in the future they may bring forward consumption decisions).

The breakdown of the Phillips curve in the late 1960s demonstrated that another 'econometric' function had been misspecified because it ignored the possibility that rising inflation might become self-fulfilling as workers and firms sought to protect their real wages and real profit margins. This meant that an **inflationary expectations** term should have been included in the equation.

It is a fact though, that the mainstream Keynesian consensus in the 1960s was abstracted from the potential instability rooted in the institutional nature of wage and price setting. Instead, policymakers pursued the attractive notion that they could permanently maintain low unemployment as long as they ensured that effective demand was sufficient, given the non-government sector's saving plans and any demand leakages from net exports.

Another reason why the discussions about instability were largely ignored is that the 'textbook' model of the Phillips curve was very attractive in its simplicity. Textbooks typically stylise discussions and eschew complicated stories for the sake of pedagogy. We have taken care to resist this tendency in this text. We consider a rich treatment of institutions and history to be an important part of learning macroeconomics.

18.3 The Accelerationist Hypothesis and the Expectations Augmented Phillips Curve

Introduction

The legacy of the earlier Keynes vs. Classical debate persisted through the 1950s and 1960s. The neoclassical school was unwilling to accept the basic insights provided by Keynes that effective demand drives output and national income and that the capitalist monetary system is susceptible to crises of overproduction and unemployment.

In the late 1960s, the ongoing debate about the effectiveness of fiscal and monetary policy in stabilising the economic cycle was revisited within the Phillips curve framework. A group of economists, centred at the University of Chicago, were opposed to government attempts to maintain full employment. Their argument largely reflected their belief that a self-regulating free market would generate optimal outcomes. Thus, they were adherents of the neoclassical model that considered most government intervention to be problematic (see the discussion of the orthodox, neoclassical approach in Chapter 1).

This debate at the microeconomic level was manifested in demands for widespread deregulation in product, labour and financial markets and a major reduction in the size of government.

At the macroeconomic level, the Phillips curve was the contested battleground. During the 'stable' Phillips curve policy era, policymakers assumed they could target a low unemployment rate and incur a modest inflation rate as a consequence. The extent to which inflation rose was determined by the slope of the Phillips curve, which was considered to be relatively low, so the Phillips curve was close to being flat.

The emerging Chicago Monetarists, who eschewed government intervention, challenged that view and asserted that there was no permanent (long-run) trade-off between inflation and unemployment. They claimed that ultimately the market would ensure unemployment stabilised at around its so-called **natural rate** and that any attempts by government to push unemployment below this rate would lead to accelerating inflation.

Chicago economist Milton Friedman was the most vocal Monetarist and in a famous article in 1968 outlined what became known as the **accelerationist hypothesis**.

Expectations of inflation

BOX 18.1 EXPECTATIONS OF INFLATION: AN HISTORICAL NOTE

The English economist Arthur Joseph Brown published *The Great Inflation* in 1955, three years before Phillips came out with his work. Brown not only foresaw the instability of the wage-unemployment relationship but also recognised that the development of a wage-price spiral is dependent on the "aims of the two parties who are competing for the real income of the country or their success in achieving those aims" (Brown, 1955: 105).

This is reminiscent of the literature we examined in Chapter 17, which depicts the inflation process as resulting from incompatible claims on total nominal income by workers, and firms.

Brown also understood in making these claims that the antagonists formed expectations of future price movements. For example, workers want to protect the real equivalent of their wage incomes and make money wage demands accordingly.

As seen in Box 18.1, the introduction of the so-called expectations augmented Phillips curve by Friedman in 1968 did not represent a sudden new discovery that expectations of inflation in wage adjustment are important. Rather, the emergence and promotion of the model can be seen as the culmination of a long campaign by Friedman and others to restore the Quantity Theory to the status it had prior to the Keynesian revolution.

Friedman and Edmund Phelps (who in 1968 also authored an important paper incorporating expectations into the Phillips curve) have thus wrongly been given the credit for the introduction of expectations into the

Phillips curve and raising the issue of its stability. However, it was the empirical instability in the relationship between unemployment and inflation that opened the way for the so-called Monetarist paradigm in macroeconomics to gain ascendancy.

The Monetarists reinterpreted the inflation/unemployment trade-off by adding the role of inflationary expectations, and in doing so, revived the Classical (pre-Keynesian) notion of a natural unemployment rate (defined as equivalent to full employment). The devastating consequence of this assertion was the rejection of a role for demand management policies to reduce unemployment to its frictional component.

The Keynesians had adopted the Phillips curve – a macroeconomic relationship – yet they had developed very little microeconomic theory to underpin it. They had justified the Phillips curve as a competitive adjustment process, such that if there was growing demand for labour, money wages rose as unemployment fell.

But Monetarists claimed that workers cared about real wages rather than nominal wages because the real wage represented a worker's capacity to buy goods and services. Thus, the original Phillips curve was defective because it only focused on the relationship between percentage changes in money wages and the unemployment rate, and ignored the influence of a changing price level on expectations about inflation.

The **accelerationist hypothesis** was advanced in 1968 by Milton Friedman before the empirical breakdown of the relationship between inflation and unemployment emerged in the early 1970s. So, while the Phillips curve presented the Monetarists with the opportunity to debate the failings of the mainstream Keynesian analysis, it was the subsequent empirical havoc created by the 1970s oil price shocks that added weight to their theoretical arguments (which were later shown to be deficient (see below)).

The misspecification of the earlier Phillips curve, which had ignored inflationary expectations, was not significant while inflation was negligible. Once inflation rates soared world wide as oil prices soared in the early 1970s, the Phillips curve relationships broke down. As a result, the Monetarist concept of a natural rate of unemployment appeared to be validated, along with the rejection of aggregate demand management through fiscal policy.

There were two basic propositions that Friedman asserted in his attack on the Phillips curve. First, he claimed that there is a **natural rate of unemployment** which is determined by the underlying structure of the labour market and the rate of capital formation and productivity growth. He believed that the economy always tends to move back to that level of unemployment even if the government attempts to use fiscal and monetary policy expansion to reduce unemployment.

He noted that the natural rate of unemployment is not "immutable and unchangeable" but is insensitive to monetary (aggregate demand) forces (Friedman, 1968: 9). In other words, he argued that increasing nominal aggregate demand would not reduce the natural rate.

He also claimed that the 'natural' base rate is both "man-made and policy-made" (Friedman, 1968: 9). For example, Monetarists argued that imposing minimum wages and providing unemployment benefits would increase the natural rate.

The concept of the natural rate of unemployment that Friedman developed follows from the analysis of the Classical labour market. We briefly summarise the arguments here, while a detailed discussion of Classical theory can be found in Chapter 11. The Monetarist approach was an attempt to restore the legitimacy of the ideas that Keynes demolished in the 1930s. Friedman asserted that real wages, rather than money wages, are the relevant object of concern from the perspective of firms and workers. Workers supply labour based on the opportunity cost of leisure, which is the income given up by an extra hour of leisure. That is, the real wage per hour. On the other hand, firms employ labour to maximise profit, so the demand for labour is also a function of the real wage.

It was argued that real wages would adjust to ensure that the labour market clears at the 'natural rate' level of unemployment (that is, the demand for labour equals the supply of labour). While there might be temporary deviations around that rate, for reasons we will explore next, the economy would always tend to move back to the natural rate of unemployment.

The natural rate was thus conceived as the level of unemployment established as a result of frictions in the labour market. Friedman considered that these frictions could include the distortions arising from policy decisions, which were noted above. The magnitude of the natural rate is not affected by the level of aggregate demand, so it has no cyclical component.

Friedman's second argument was that the Phillips curve is, at best, a short-run relationship that can only be exploited through aggregate demand expansion as long as workers suffer from money illusion by confusing money wage rises with real wage rises. In other words, the validity of any given short-run Phillips curve is dependent on workers assuming that the prevailing rate of price inflation is stable.

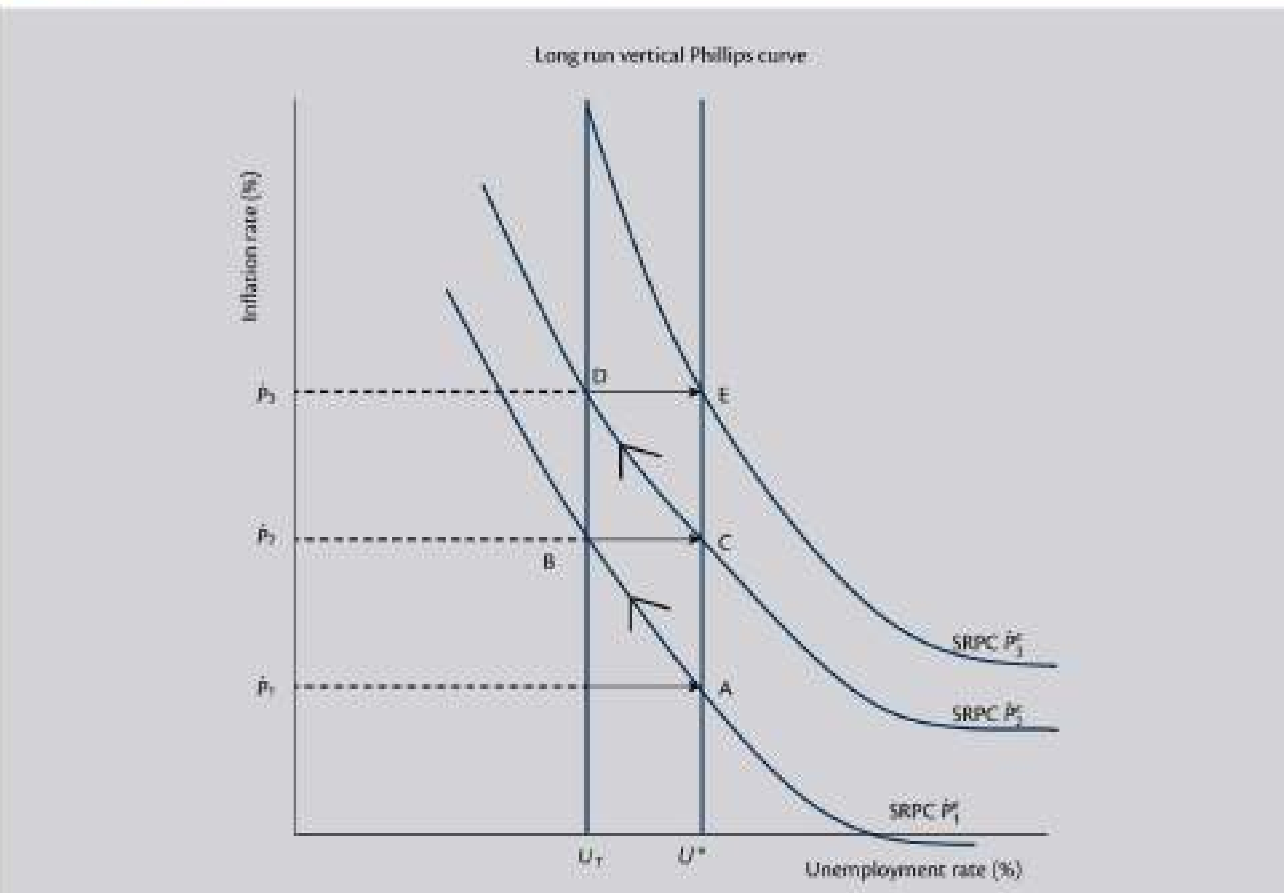
However, Friedman and others argued that eventually workers would realise that their real wage was being eroded as price inflation outstripped money wages growth in the light of the expansion of aggregate demand. Thus, they would form expectations of higher inflation. As a consequence, workers would build these inflationary expectations into their future outlook and pursue money wage increases which reflected not only the state of the labour market (relative strength of demand and supply) but also how much they expected prices to rise in the period.

The Monetarists argued that if the government attempted to reduce unemployment below the natural rate, then as the inflation rate rose, workers would demand even higher money wages growth to achieve their desired real wage levels. Ultimately, this would result in a rising rate of inflation.

Figure 18.4 captures this **accelerationist** hypothesis. The short run Phillips curves are shown as conditional on a specific expectation of inflation held by the workers. The superscript *e* denotes **expected inflation**. We use the terminology, 'expectations are realised' to denote a state where inflation expectations are equal to the actual inflation outcome.

We start at Point A, where the inflation rate is $\dot{P}_1$ and the unemployment rate is at its so-called 'natural rate' (U^*). At this point, the expectations about the rate of inflation held by workers ($\dot{P}_1^e$) are consistent with the actual inflation rate of $\dot{P}_1$. According to Friedman, the labour market would be operating at the natural rate of unemployment whenever inflationary expectations are realised.

Figure 18.4 The expectations augmented long run Phillips curve



To see how the accelerationist hypothesis plays out, we assume that the government is under political pressure and concludes that the unemployment rate, U^*, is too high. It believes it can use expansionary fiscal and monetary policy to target a lower rate of unemployment, U_1. It also thinks that it can exploit the Phillips curve trade-off and move the economy to Point B, with a higher inflation rate ($\dot{P}_2$) as the cost of lower unemployment. Consequently, the government stimulates nominal aggregate demand to push the economy to point B. The increased demand for labour pushes up the inflation rate (to $\dot{P}_2$) and money wage rates also rise in the labour market. The accelerationist hypothesis assumes that the price level accelerates more quickly than money wages and as a consequence the real wage falls.

The Monetarists resurrected the Classical labour market and placed it at the centre of their attack on Keynesian macroeconomics. Accordingly, firms will offer more employment because the real wage has now fallen and the demand for labour is an inverse function of the real wage.

Why would workers supply more labour if the real wage was falling? In the Classical labour market it is assumed that labour supply is a positive function of the real wage, so workers will withdraw their labour if the real wage falls.

The Monetarist approach overcomes that apparent problem by imposing different expectations on workers and firms. Firms are assumed to always have perfect information about prices and wages so they know the level of the actual real wage at any point in time. However, the workers are assumed to gather information about the inflation rate in a lagged or adaptive fashion and thus could be fooled into believing that the real wage is rising (because their money wages are rising), when in fact, it is falling. This is known as asymmetric information. Thus, workers are assumed to be initially oblivious to the higher inflation, that is, their inflationary expectations do not adjust to the actual inflation rate immediately. As a consequence, they mistake the rising nominal wages for an increasing real wage and willingly supply more labour even though the real wage has actually fallen.

The central proposition of the Classical labour market is that workers care about real wages not money wages. The accelerationist hypothesis added the idea that workers form **adaptive expectations of inflation**, which means that it takes some time for them to differentiate between movements in money wages and movements in real wages.

Monetarists asserted that Point B of figure 18.4 is unstable and can only persist as long as workers are fooled into believing that the money wage increases they received are equivalent to real wage increases.

But inflationary expectations adapt to the actual higher inflation rate after a time. Once workers increase their inflationary expectations to $\dot{P}_2$, then the **short run Phillips curve** (SRPC) shifts out because their expectations of the underlying rate of price inflation have risen. The labour market will then settle at Point C, which is consistent with the new (higher) expected inflation rate.

The path the labour market takes as inflationary expectations adjust to the actual inflation rate and the SRPC shifts (that is, from Point B to Point C, or from Point D to Point E when the expected inflation rate has shifted to $\dot{P}_3$) is an empirical matter. But for Monetarists, once inflationary expectations have fully adjusted to the current inflation rate (at Points C and E, for example), the economy will return to the natural rate of unemployment (U^*).

For Friedman, the short-run dynamics of the labour market were driven by the capacity of the government to 'fool' workers into believing that the inflation rate is lower than the actual inflation rate. As long as the actual rate of inflation is underestimated by workers, the government can maintain the unemployment rate below the natural rate, but at a cost of rising inflation.

This narrative seeks to explain mass unemployment in the same way. This would occur when the economy is operating with the unemployment rate to the right of U^* in Figure 18.4. The Friedman interpretation is that mass unemployment occurs when workers think the real wage implied by a nominal wage offer is too low, because they wrongly believe that inflation is higher than it is. We slide down the short run Phillips curve with unemployment rising above the natural rate, actual inflation below expected inflation, and workers believing that the real wage offers are too low to accept. As a consequence, they start quitting their jobs and/or refuse to take new job offers, thinking it is better to search for positions which offer higher real wages. Once they realise they have mistakenly thought inflation was higher than it actually is, they start to accept the job offers at the current money wage levels, thus increasing the labour supply and the economy moves back up a short run Phillips curve toward the natural rate of unemployment, U^*.

Friedman was thus forced to explain changes in unemployment in terms of swings in the supply of labour that are driven by misconceptions of the actual inflation rate. At the empirical level this theory predicts that quit rates will fall as employment rises.

If the quit rate is indeed **countercyclical**, the resulting changes in labour supply will be consistent with Friedman's theory. However, the empirical evidence is that quit rates are **pro-cyclical**, which means that they rise when the labour market is strong and workers feel confident about their chances of securing new jobs (after quitting their current jobs) and fall when the labour market is weak and workers fear ongoing unemployment. This is exactly the opposite of what would be required to substantiate Friedman's natural rate theory.

American economist, Lester Thurow summarised this issue succinctly: "why do quits rise in booms and fall in recessions? If recessions are due to informational mistakes, quits should rise in recessions and fall in booms, just the reverse of what happens in the real world." (Thurow, 1983: 185)

The introduction of the role of inflationary expectations in the Phillips curve focused attention on how such expectations are formed. What behavioural models could be invoked to capture expectations? There were two main theories advanced by economists: (a) adaptive expectations, and later (b) rational expectations. Both theories considered the formation of expectations to be endogenous to the economic system. That is, developments within the system determine how workers (and firms) foresee the future course of inflation.

The algebra of the expectations augmented Phillips curve

Here we present a more analytical version of the Friedman natural rate hypothesis. The original Phillips curve related the growth of money wages to the unemployment rate. Friedman claimed that the simple version of the Phillips curve, whether specified in its original form or in the price inflation form, overlooked the fact that workers would be concerned about the growth in real wages. In other words, the rate of money wages growth would be influenced by the expected inflation rate, independently of the state of the labour market.

This conjecture led Friedman to incorporate a term for the influence of inflationary expectations in the wage bargaining process in the wage Phillips curve:

$$(18.5) \qquad \dot{W}_t = \alpha - \beta U_t + \varphi \dot{P}_t^e \qquad 0 \leq \varphi \leq 1$$

The additional term $\dot{P}_t^e$ represents inflationary expectations that are formed by workers, which condition the wage bargaining process. We assume that the coefficient φ lies between zero and one. If $\varphi = 0$, then wage inflation will only depend on the state of the labour market captured by the excess demand term (βU_t). If $\varphi = 1$, then any change in inflationary expectations ($\dot{P}_t^e$) is passed on fully to wages growth.

We assume that workers' expectations of inflation in period t are based on price information from the previous period ($t - 1$). They then bargain for wages growth in period t based on these inflationary expectations.

The **expectations augmented Phillips Curve (EAPC)** can be written as:

$$(18.6) \qquad \dot{P}_t = \alpha_0 - \beta U_t + \varphi \dot{P}_t^e$$

where again the constant term is $\alpha = \alpha_0 + \dot{\gamma}$.

In terms of Figure 18.4, the inflationary expectations term on the right-hand side of Equation (18.6) ($\varphi \dot{P}_t^e$) shifts the **short run Phillips curve**, $\dot{P}_t = \alpha - \beta U_t$.

If workers' inflationary expectations increase, then the short run Phillips curve shifts out, and vice versa.

After the EAPC replaced the simple Phillips curve as the main framework for considering the relationship between inflation and unemployment, economists began to focus on the value of φ. Many econometric studies were conducted to estimate its value.

Why does the value of φ matter?

Friedman defined the long-run steady state (stable) inflation rate to be when workers' inflationary expectations equalled the actual rate of inflation. At this point, he claimed the economy would be operating at the natural rate of unemployment.

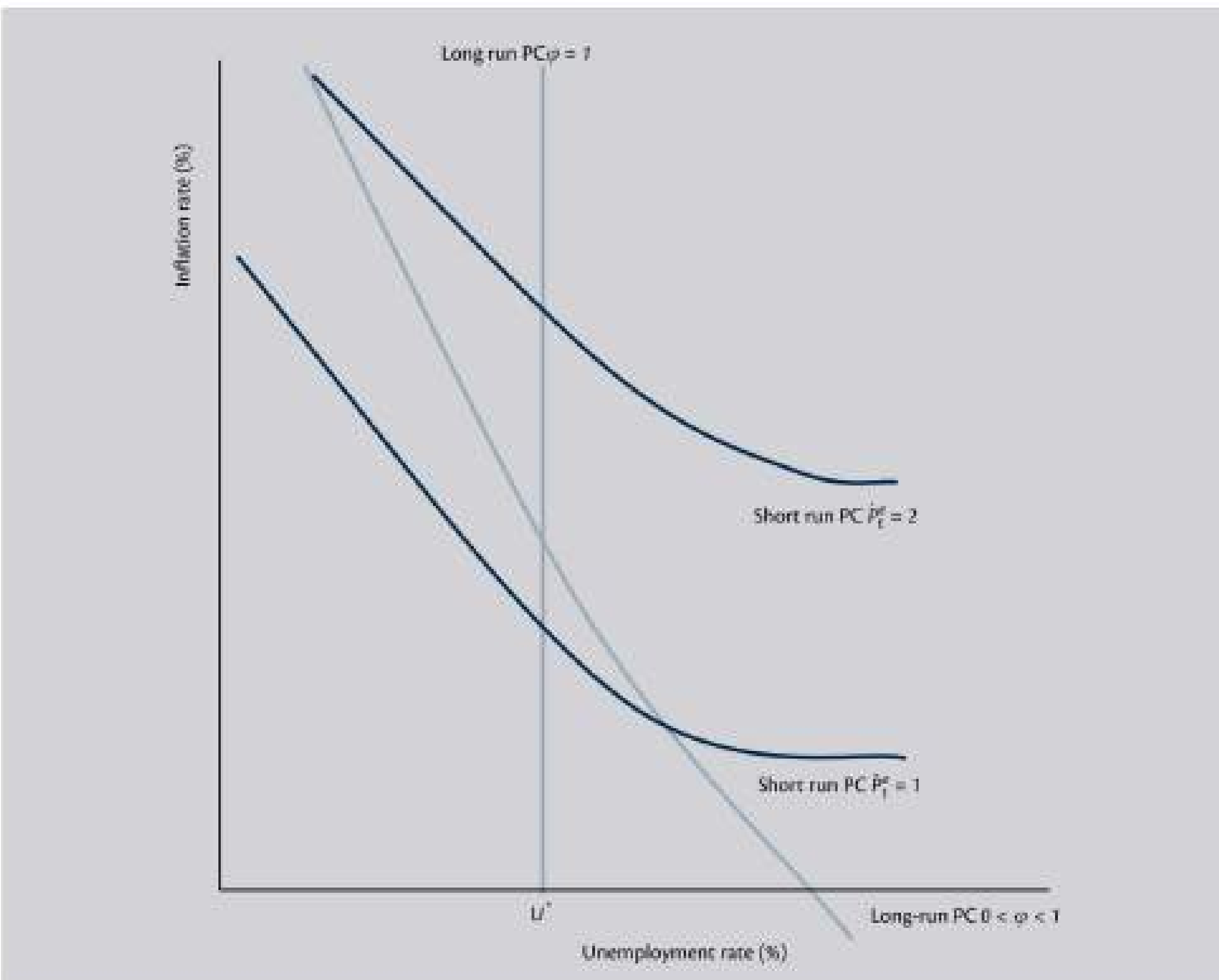
In that case, the EAPC would collapse to what is referred to as the **long-run steady state Phillips curve**. We replace $\dot{P}^e_t$ by $\dot{P}_t$ in Equation (18.6) and rearrange the equation to give:

$$\dot{P}_t(1-\varphi) = \alpha - \beta U_t \tag{18.7}$$

$$\dot{P}_t = \alpha/(1-\varphi) - \beta U_t/(1-\varphi)$$

This relationship should be examined carefully because it looks similar to the short run Phillips curve (Equation 18.4) except the coefficients are now divided by the term $(1 - \varphi)$. The negative slope of this long run Phillips curve, $-\beta/(1 - \varphi)$, is steeper than the slope of the short run Phillips curve, $-\beta$. The closer φ is to one, the steeper is the slope of the long run Phillips curve. Once φ equals one, the slope becomes vertical and there is no longer any relationship between inflation and the unemployment rate. In other words, the trade-off vanishes.

Figure 18.5 Short and long run Phillips curves



Notes: $\dot{P}^e_t$ is the expected inflation rate at time t. The higher the expected rate of inflation, the greater is the intercept of the short run Phillips curve with the long-run PC.

Figure 18.5 depicts the two cases. There is a family of short run Phillips curves (SRPC) (two are shown). The Phillips curve drawn on the assumption that $0 < \varphi < 1$ is steeper than the short run curves but non-vertical. It means that in the long run there is still a trade-off between the inflation rate and the unemployment rate, but it is a steeper trade-off than occurs in the short run before inflationary expectations adjust upwards to the new inflation rate.

The long run Phillips curve for $\varphi = 1$ is vertical. This means there is no long-run trade-off between inflation and the unemployment rate that can be exploited by the government. Under these assumptions, the economy always tends back to the natural rate of unemployment U^* once inflationary expectations have adjusted to the actual inflation rate (see below).

You can now see why the economists embraced this flawed framework were interested in the value of φ. For Keynesians, a value of φ of less than unity maintained their policy position that the government could use expansionary fiscal and monetary policy to reduce the unemployment rate should they consider the current rate to be too high.

For Monetarists, a value of $\varphi = 1$, was consistent with their claims that the Keynesian aggregate demand management framework was flawed and would cause rising inflation should the government try to push the unemployment rate below its natural rate, which is based on inflationary expectations being equal to the actual inflation rate.

To see how the natural rate of unemployment emerges out of this framework, we can solve Equation (18.7) for the long-run unemployment rate. After the relevant algebraic manipulation, we get:

$$(18.8) \qquad U^* = \alpha/\beta - \dot{P}_t(1-\varphi)/\beta$$

which shows there is still a trade-off in the long run between unemployment and inflation as long as $\varphi \neq 1$. Once $\varphi = 1$, the long-run unemployment rate becomes Friedman's natural rate, and the equation representing that case is written as:

$$(18.9) \qquad U^* = \alpha/\beta$$

This means that in the **Friedman natural rate hypothesis** there are only two specific factors which influence the long-run or natural rate of unemployment: (a) the rate of growth of productivity which is captured in the α term; and (b) the short run responsiveness of wage inflation to movements in the unemployment rate (β). Note that given β is assumed to be positive, the term (α/β) is positive.

As a result, the higher is the growth in productivity, other things being equal, the lower will be the natural rate. The Monetarists assumed that productivity growth was a structural phenomenon and invariant to aggregate demand policies.

It is clear that in the expectations augmented Phillips curve framework, the government could only achieve temporary reductions in the unemployment rate below the natural rate as long as it could maintain a wedge between the expected inflation rate and the actual inflation rate. Once the workers' inflationary expectations adjusted, then the trade-off disappeared, and the economy would return to the natural rate of unemployment, albeit with higher inflation. Continued attempts at driving down the unemployment rate below the natural rate would, according to the Monetarists, just result in accelerating inflation.

Specification of inflationary expectations

Adaptive expectations hypothesis

The assumption that workers form their expectations of inflation in an adaptive manner allowed the Monetarists to conclude that government attempts to reduce the unemployment rate would only cause accelerating inflation and that the economy would always tend back to the natural rate of unemployment.

The only way the government could sustain an unemployment rate below the natural rate using aggregate demand stimulus would be if they continually drove the rate of price inflation ahead of money wage inflation and workers continually misperceived the true (price) inflation rate.

The adaptive expectations hypothesis is expressed in terms of the past history of the inflation rate, assuming that workers learn from their past forecasting errors, and adapt their price inflation expectations.

The following model expresses this idea:

$$(18.10) \quad \dot{P}^e_{t+1} = \dot{P}^e_t + \lambda(\dot{P}_t - \dot{P}^e_t) \quad 0 < \lambda < 1$$

The left-hand side of Equation (18.10) is the expected inflation rate in the next period ($t + 1$) and is based on the experiences of workers in period t.

Equation (18.10) has two components on the right-hand side.

- $\dot{P}^e_t$ is the expected inflation rate in the current period. Thus, workers use this inflation rate as a baseline to what they think the inflation rate in the next period will be.
- The term $\lambda(\dot{P}_t - \dot{P}^e_t)$ captures the **forecast error** in the current period. $\dot{P}^e_t$ is the expectation that workers formed in period $t - 1$ of the inflation rate in period t. The difference between that expectation and the actual rate that occurred is the size of their forecast error. The coefficient λ measures the strength of adaptation to error. The higher is λ, the more responsive workers' expectations will be to the actual rate of inflation. If $\lambda = 1$, then the expected inflation rate in period $t + 1$ is simply the actual rate of inflation in period t.

Note that even if expectations quickly adapt to errors ($\lambda = 1$), if inflation is rising, workers will still make errors in their forecasts year after year. As we will see, this led some economists to reject the adaptive expectations hypothesis that Friedman had proposed.

The rational expectations hypothesis

An extreme form of Monetarism, which became known as **New Classical Economics**, posits that no policy intervention from government can be successful because so-called economic agents (for example, households and firms) form expectations in a rational manner.

This literature, which evolved in the late 1970s, claimed that the stimulation of aggregate demand, say via fiscal policy, would not only be ineffective in real terms but also highly inflationary.

The theory claimed that since economic agents form their expectations rationally, they would be able to anticipate any government policy action and its intended outcome and change their behaviour, which in turn would undermine the desired impact of the policy.

For example, households might anticipate a rise in government spending and predict that taxes would rise in the future to pay back the deficit. As a result, households would reduce their own spending to save for the higher taxes. That action would thwart the expansionary impact of the increase in public spending.[1]

In this context, Monetarists like Milton Friedman claimed that the government could exploit a short run Phillips curve for a time with expansionary policy by tricking workers into thinking their real wages had risen when in fact their money wage increases were lagging behind the inflation rate. The unanticipated inflation would then induce the workers to supply a higher quantity of labour than would be forthcoming at the so-called natural rate of output (defined in terms of a natural rate of unemployment).

Recall that under adaptive expectations, economic agents are playing catch-up all the time; so workers' expectations of inflation take time to catch up with the actual inflation rate. Once they adjusted to the actual inflation rate and realised that their real wage had actually fallen, they would withdraw their labour so that the natural level of output would be restored.

The use of adaptive expectations to represent the way workers adjusted to changing circumstances was criticised because it implied an implausible irrationality. In a period of continually rising prices, workers' expectations would never catch up. Why wouldn't they realise after a few periods of errors that they were systematically underforecasting and seek to compensate by overshooting the next period?

The theory of rational expectations was developed in part, to meet these objections. When forming their expectations, economic agents were considered to act in a rational manner consistent with the assumptions in mainstream microeconomics pertaining to *Homo economicus*.[2] This assumes that economic agents used all the information that was available and relevant at the time when forming their views of the future.

What information *do* they possess? The rational expectations (RATEX) hypothesis claims that individuals essentially know the true economic model that is driving economic outcomes and can make accurate predictions of these outcomes. Any forecasting errors are random. The proponents of RATEX said that predictions derived from rational expectations are on average accurate.

These proponents assumed that all people understand the economic models that policymakers use to formulate their policy interventions. The most uneducated person is therefore assumed to have a highly sophisticated grasp of the structural specification of the economy that treasury and central banks deploy in their policymaking processes.

Further, the lay public is assumed to be able to employ this impressive knowledge to perfectly predict how policymakers will respond (in both direction and quantum) to past policy forecast errors. According to the RATEX hypothesis, people are able to anticipate both policy changes and their impacts.

As a result, any 'pre-announced' policy expansions or contractions will have no effect on the real economy. For example, if the government announces it will be expanding the deficit and adding new base money, we will also assume immediately that it will be inflationary and will not alter our real demands or supply (so real outcomes remain fixed). Our response will be to simply increase the value of all nominal contracts and thus generate the inflation that we predict via our expectations.

Further, as rational agents understand the economic models used by policymakers, they can predict policy even if it is not announced. Rational agents cannot be surprised by predictable policy, and any policy formulated on the basis of available data and models used by policymakers will be foreseen by rational agents.

The only exception would be if policymakers formulated policy in an unpredictable manner, one that was not based on data and models. Since rational agents are supposed to have at hand the same data and models as those used by policymakers, for policy to be unpredictable it would have to be formulated in a random manner. For example, monetary policymakers would use a random number generator to determine where to set money supply growth rate targets or interest rate targets. Fiscal policymakers would use a random number generator to determine where to set government spending for the next year. Only in such cases would government be able to 'fool' rational agents, but such policy would be silly.

The government can thus never trick the private sector so long as policy is predictable. The introduction of rational expectations into the debate thus went a step further than the Monetarists who conceded that governments could shift the economy from the 'natural level' by introducing adaptive expectations. With rational expectations, this sort of fooling is eliminated.

The New Classical economists denied that governments could alter the course of the real economy at all. In other words, there was not even the possibility of a short-run trade-off between inflation and the unemployment rate. Workers would always know the future inflation rate and would build it fully into each round of money wage bargaining. The economy would thus always stay on the long run Phillips curve.

While there are some very sophisticated theoretical critiques of the RATEX hypothesis (for example, the Sonnenschein-Mantel-Debreu theorem) which extend the notion of the fallacy of composition, some simple reflection suggests that the informational requirements necessary for the hypothesis to be valid are beyond the scope of individuals.

A relatively new field of study called behavioural economics has attempted to examine how people make decisions and form views about the future. The starting point is that individuals have what are known as cognitive biases (as mentioned in Chapter 8), which constrain their capacity to make rational decisions.

The Global Financial Crisis (GFC) was not the first time that models employing rational expectations categorically failed to predict major events. RATEX-based models have failed to account for even the most elemental macroeconomic outcomes over the last several decades. They categorically fail to predict movements in financial, currency and commodity markets. Rational expectations impose a mechanical forecasting rule onto individual decision making when in fact these individuals exist in an environment of endemic uncertainty in which the future is unknowable.

As we will see in later chapters, endemic uncertainty is a major problem facing decision makers of all types in a capitalist monetary economy. Uncertainty about economic events, such as movements in asset prices or job security, may encourage individuals to hold money, which is the most liquid of all assets and is a store of value.

In the real world of course, people have imperfect knowledge of what information is necessary for forecasting and even less knowledge of how their choices – made on the basis of this imperfect or incomplete information – might impact future outcomes. We also do not know how we will react to changing circumstances until we are confronted with them. The nature of endemic uncertainty is that we cannot know the full range of options that might be presented to us at some time in the future.

18.4 Hysteresis and the Phillips Curve Trade-off

While economists focused on how fast individual expectations might respond to rising inflation and the role that inflationary expectations play in wage and price formation, a new strand of literature emerged which challenged the Monetarist contention that there is no long-run trade-off between inflation and the unemployment rate.

At the empirical level it was noted that the estimates of the unobserved natural unemployment rate (sometimes called the **non-accelerating inflation rate of unemployment** or **NAIRU**) which were derived from econometric models seemed to track the actual unemployment rate with a lag.

At the time that Monetarism became influential in economic policymaking, unemployment rates were rising due to the policy response to the major oil price rises in the early and mid-1970s that caused accelerating inflation. We will examine this period separately later. The estimates of the natural rate seemed to rise too, and without any consistent explanation.

As an example, in Figure 18.6, we show Australian Treasury and OECD estimates of the NAIRU and the corresponding unemployment rate for Australia from 1960 to 2015. The Treasury estimates end in 2011 while the OECD estimates only begin in 1978.

First, why did the Treasury estimates of the NAIRU, which are meant to reflect 'structural' factors jump so violently in 1974, around the same time the actual unemployment rate rose sharply? This period was highly turbulent (OPEC oil crisis) and marked the end of the post-war full employment era when unemployment rates were usually below two per cent. You can clearly see that the estimated NAIRU tracks the actual unemployment rate upwards. No coherent explanation has ever been given to explain that jump. Structural factors tend to impact slowly and gradually.

Second, why are the Treasury estimates flat after 1974 and quite different to the OECD estimates, which track the actual unemployment rate?

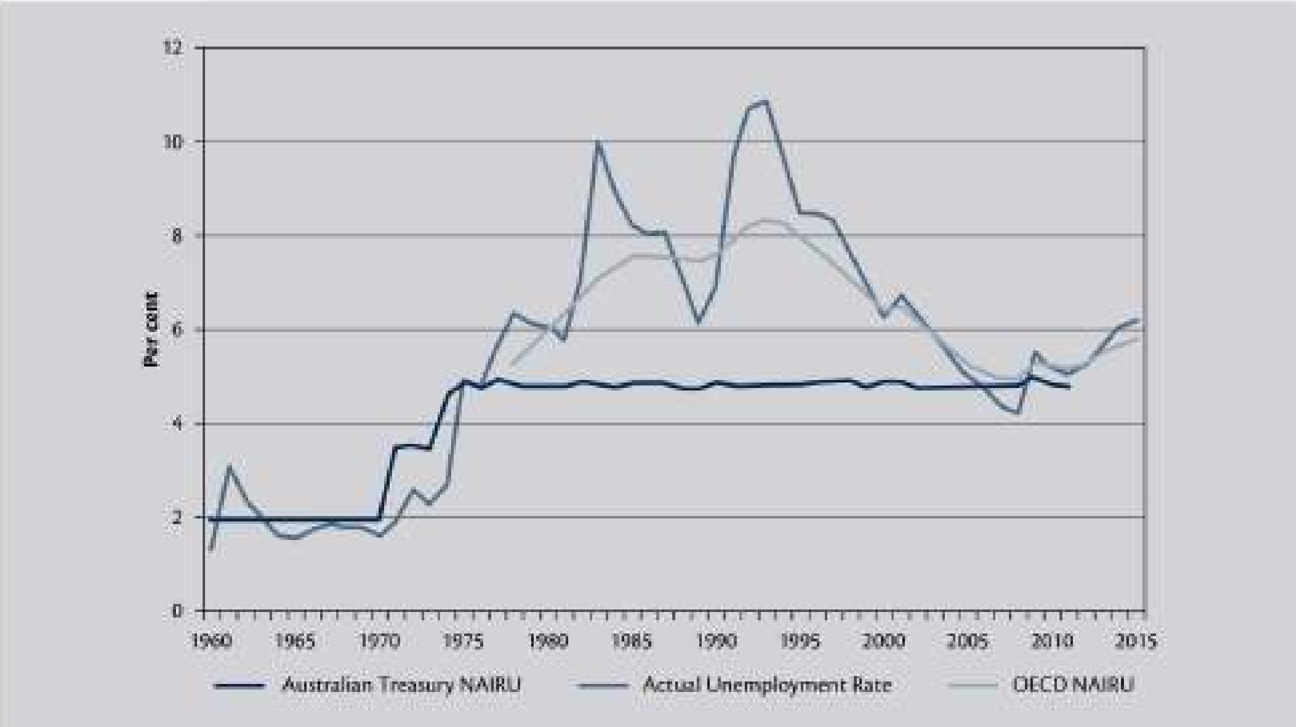
Third, with the exception of 2001, the OECD estimates of the NAIRU steadily declined from 1993 to 2007. Likewise, there was a steady decline in OECD estimates of the NAIRU for the UK and USA for more than ten years from the early 1990s.

These observations led economists to question the idea that there is a **cyclically invariant** natural rate of unemployment. It appeared that the best estimates of an unemployment rate that was consistent with stable inflation at any point in time were highly cyclical, since they followed the actual unemployment rate, which reflects the business cycle.

While there are various explanations to rationalise the way the estimated natural rates of unemployment fell over the 1990s (for example, demographic changes in the labour market with the youth cohort's share of the labour force declining), one plausible explanation is that there is no separate informational content in these estimates. They just reflect in some lagged fashion the dynamics of the unemployment rate, that is, the **hysteresis hypothesis**.

This was a new theory that emerged to explain the apparently cyclical relationship between the equilibrium unemployment rate and the actual unemployment rate. The early work showed that the increasing NAIRU estimates (based on econometric models) merely reflected the decade or more of high actual unemployment rates and restrictive fiscal and monetary policies. Thus, these estimates were not indicative of increasing structural impediments in the labour market that were due to, for example, demographic changes, which could result in an influx of young unskilled workers into the labour market or rising minimum wages and/or welfare distortions, such as more generous unemployment benefits.

Figure 18.6 Annual Australian unemployment rate, Treasury and OECD NAIRU estimates, 1960–2015



Source: Authors' own. Data from Australian Bureau of Statistics.

The **hysteresis effect** describes the interaction between the actual and equilibrium unemployment rates. The significance of hysteresis is that the equilibrium unemployment rate is associated with stable prices, and at any point in time should not be conceived of as a rigid non-inflationary constraint on expansionary macro policy. The equilibrium rate itself can be reduced by policies, which reduce the actual unemployment rate.

The importance of hysteresis is that a long-run inflation-unemployment rate trade-off can still be exploited by the government, and this would invalidate one of the major planks of Monetarism.

One way to explain hysteresis is to focus on the way in which the labour market adjusts to cyclical changes in economic activity. Recessions cause unemployment to rise and due to their prolonged nature, the short-term joblessness becomes entrenched in long-term unemployment. The unemployment rate behaves asymmetrically with respect to the business cycle, which means that it jumps up quickly but takes a long time to fall again.

The labour market adjustments that accompany an economy with lots of slack, which could lead to hysteresis, are well documented. Training opportunities are provided with entry level jobs and so the (average) skill of the labour force declines as vacancies fall.

New entrants to the labour force who cannot find jobs are denied relevant skills (and socialisation associated with stable work patterns) and unemployed workers face skill obsolescence. Both groups need jobs in order to update and/or acquire relevant skills. Skill (experience) upgrading also occurs through mobility, which is restricted during a downturn.

The idea is that structural imbalance increases in a recession due to the cyclical labour market adjustments commonly observed in downturns, and decreases at higher levels of demand as the adjustments are reversed. Structural imbalance refers to the inability of the actual unemployed to present themselves as an effective excess supply. The longer that people remain unemployed, the harder it becomes to convince potential employers that they have the requisite skills and the appropriate attitudes.

The algebra of hysteresis

Here we will learn that if there is hysteresis present in the labour market, then a long-run trade-off between inflation and the unemployment rate is possible even if the coefficient in the augmented term in the Phillips curve (φ) is equal to unity. This result was shown in Mitchell (1987).

At any point in time there might be an equilibrium unemployment rate which is associated with wage stability in that it temporarily constrains the wage demands of the employed and balances the competing distributional claims on output. We might call this unemployment rate the **macroequilibrium rate of unemployment** (**MRU**). Abstracting away from other causes of inflation (such as pressures arising from rising costs of resources and other inputs to production, as well as attempts by firms to raise their mark-ups over costs to increase profits), the MRU will be consistent with price stability.

The interaction between the actual and equilibrium unemployment rates has been termed the hysteresis effect. The significance of hysteresis – if it exists – is that the unemployment rate associated with stable prices at any point in time should not be conceived of as a rigid non-inflationary constraint on expansionary macro policy.

The equilibrium rate itself can be reduced by policies, which reduce the actual unemployment rate. Thus, we use the term MRU as the non-inflationary unemployment rate (as distinct from the Monetarist concept of the NAIRU) to highlight the hysteresis mechanism, which is driven by the business cycle.

The idea is that structural imbalance increases in a recession due to the cyclical labour market adjustments commonly observed in downturns, and decreases at higher levels of demand as the adjustments are reversed. Structural imbalance refers to the inability of the actual unemployed to present themselves as an effective excess supply.

To see how hysteresis alters the Phillips curve, we start with a standard wage inflation equation such as:

$$\dot{W}_t = \alpha - \beta(U_t - U_t^*) + \varphi\dot{P}_t^e \qquad (18.11)$$

so that the rate of growth in money wages in time t, $\dot{W}_t$, is equal to a constant (α) less the deviation of the unemployment rate from its steady state value. The gap $(U_t - U_t^*)$ is just a different way of capturing the excess demand in the labour market. If the gap is positive then the actual unemployment is above the MRU and there should be downward pressure on money wage demands, other things being equal. If the gap is negative then the actual unemployment is below the MRU and there should be upward pressure on money wage demands, other things being equal. The additional term captures inflationary expectations as explained in our derivation of the EAPC.

This **hysteresis effect** – the tracking of the actual unemployment rate by the equilibrium rate of unemployment – could be represented in a number of ways. In this example, we follow Mitchell (1987) who represented U^* as a weighted average of the actual unemployment rate and the equilibrium rate in the last period.

The following model shows that the MRU adjusts to the actual unemployment with a lag:

$$(U_t^* - U_{t-1}^*) = \mu(U_{t-1} - U_{t-1}^*) \quad 0 < \mu \leq 1 \qquad (18.12)$$

This says that the **current MRU** U_t^* is equal to its value in the previous period U_{t-1}^* plus some fraction of the gap between the actual and MRU in the previous period.

The value of μ measures the sensitivity of MRU to the current state of activity. If $\mu = 0$, then $U_t^* = U_{t-1}^*$ and the MRU is invariant to fluctuations in the unemployment rate. Thus, the MRU can be identified with the NAIRU. If $\mu = 1$, then $U_t^* = U_{t-1}$, so the MRU tracks the actual unemployment rate with a one-period lag.

Substituting for the bracketed unemployment term in (18.13) yields the following wage inflation equation (Mitchell, 1987):

$$\dot{W}_t = \alpha - (\frac{\beta}{\mu})(U_{t+1}^* - U_t^*) + \varphi\dot{P}_t^e \qquad (18.13)$$

If the operation of the business cycle leads to a fall in the unemployment rate, this will lead to a fall in the MRU and over time the difference between consecutive MRUs $(U_{t+1}^* - U_t^*)$ will converge to zero (see Equation (18.12)). Thus, in contrast to the EAPC, where a rate of unemployment persistently below the NAIRU continues to add to inflationary pressure (i.e. the rate of inflation is accelerating), under the MRU model, the initial impulse to inflation from a lower unemployment rate dies away with the convergence of the MRU to the lower rate of unemployment. Thus, a lower rate of unemployment is associated with a higher but stable rate of inflation. There is a trade-off: the higher is μ, other things being equal, the flatter will be the trade-off between inflation and unemployment.

18.5 Underemployment and the Phillips Curve

As we saw in Chapter 5, underemployment has become an increasingly significant component of labour underutilisation in many nations over the last two decades. In some nations, such as Australia, the rise in underemployment has outstripped the fall in official unemployment. National statistical agencies have responded to these trends by publishing more regular updates of underemployment. They have also constructed new data series to provide broader measures of labour wastage (for example, the Australia Bureau of Statistics Broad Labour Underutilisation series, which is published on a quarterly basis).

After the major recession that beset many nations in the early 1990s, unemployment fell as growth gathered pace. At the same time, inflation also moderated and this led economists to increasingly question the practical utility of the concept of a cyclically invariant natural rate (or NAIRU) for policy purposes, quite apart from the conceptual disagreements.

This scepticism was reinforced because various agencies produced estimates of the natural rate of unemployment that declined steadily throughout the 1990s as the unemployment rate fell (see Figure 18.7). As the unemployment rate went below each natural rate estimate (and inflation continued to fall) new estimates of the natural rate were produced, which showed it had fallen further. This reinforces our conclusion that NAIRU estimates have no predictive capacity in relation to the movements between the unemployment rate and the inflation rate.

The question then arises as to why the unemployment rate and the inflation rate both fell in many nations during the 1990s. Also, what does this mean for the Phillips curve?

To understand this more fully, economists started to focus on the concept of the excess supply of labour, which is a key variable constraining wage and price changes in the Phillips curve framework.

The standard Phillips curve approach predicts a statistically significant, negative coefficient on the official unemployment rate (a proxy for excess demand). However, the hysteresis model suggests that state dependence is positively related to unemployment duration and at some point the long-term unemployed cease to exert any threat to those currently employed.

Consequently, they do not discipline the wage demands of those in work and do not influence inflation. The hidden unemployed are even more distant from the wage setting process. So, we might expect that short-term unemployment is a better excess demand proxy in the inflation adjustment function than is the overall unemployment rate (that includes long-term unemployed).

While the short-term unemployed may be proximate enough to the wage setting process to influence price movements, there is another significant and even more proximate source of surplus labour available to employees to condition wage bargaining: the underemployed.

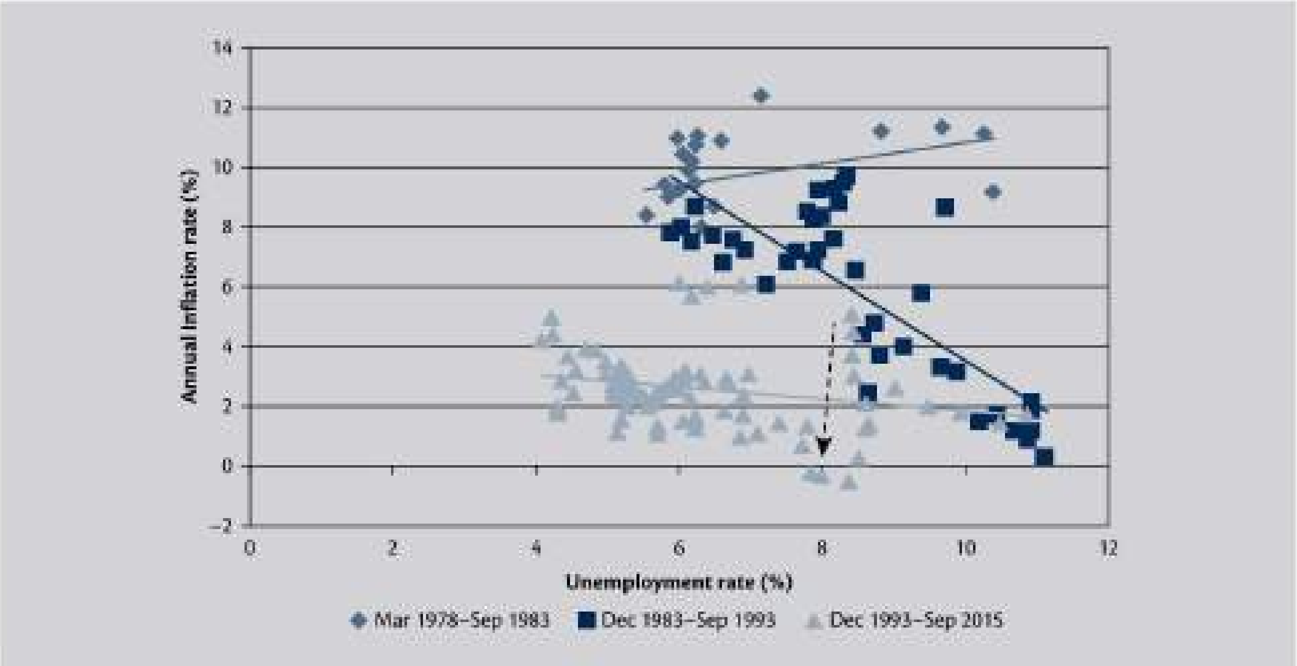
The underemployed represent an untapped pool of additional working hours that could be redistributed among a smaller pool of persons in a relatively costless fashion if employers so desired.

It is thus reasonable to hypothesise that the underemployed pose a viable threat to those in full-time work who might be better placed to set the wage norms in the economy. This argument is consistent with research in the literature that shows that wage determination is dominated by insiders (the employed) who set up barriers to isolate themselves from the threat of unemployment. Phillips curve studies have found within-firm excess demand for labour variables (such as the rate of capacity utilisation or rate of overtime) to be more significant in disciplining the wage determination process than external excess demand proxies, such as the unemployment rate.

It is plausible that while the short-term unemployed still pose a more latent threat than the long-term unemployed, the underemployed are also likely to be considered an effective surplus labour pool. In that case we might expect downward pressure on price inflation to emerge from both these sources of excess labour.

As an example, Figure 18.7 shows the relationship between the unemployment rate and inflation in Australia between 1978 and 2015. The sample is split into three subsamples. The first from March 1978 to September 1983 is defined by the starting point of the most recent consistent Labour Force data series (February 1978) and the peak unemployment rate from the 1982 recession (September 1983).

Figure 18.7 Inflation and unemployment, Australia, quarterly data, 1978–2015



Source: Authors' own. Data from Australian Bureau of Statistics, Consumer Price Index and Labour Force.

The second period, December 1983 to September 1993, depicts the recovery phase in the 1980s and then the period to the unemployment peak that followed the 1991 recession. The final period goes from December 1993 to September 2015. The solid lines are simple linear trend regressions.

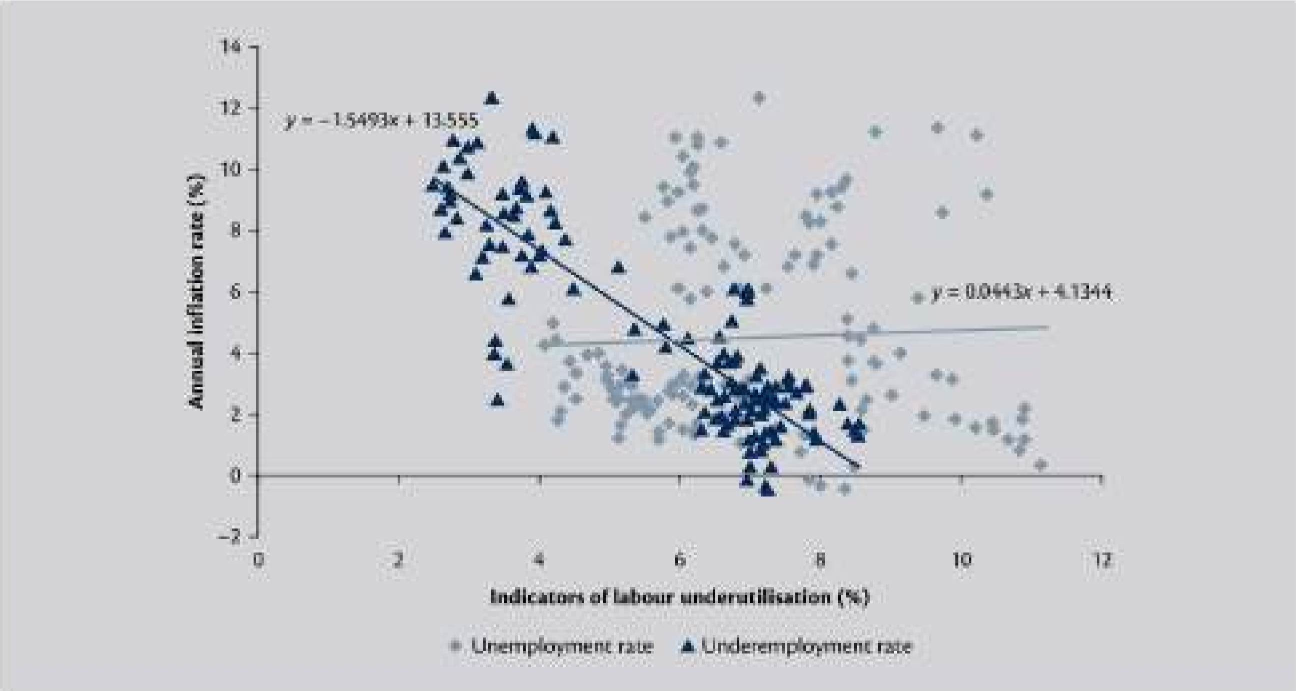
The relationship between the annual inflation rate and the unemployment rate clearly shifted after the 1991 recession. Focusing on the dashed arrow line (joining September 1995 and September 1997), this was a period when the Phillips curve began to flatten and move inwards. Over these years, the unemployment rate was stuck due to a lack of aggregate demand growth but the inflation rate was falling. This has been explained in part by the fall in inflationary expectations. The 1991 recession was particularly severe and led to a sharp drop in the annual inflation rate, and with it a decline in survey-based inflationary expectations.

The other major labour market development that arose during the 1991 recession was the sharp increase and then persistence of high underemployment. Firms shed full-time jobs, and as the recovery got under way, began to replace the full-time jobs that were shed with part-time opportunities. Even though employment growth gathered pace in the late 1990s, a majority of those jobs in Australia were part time. Further, the part-time jobs were increasingly of a casual nature.

Figure 18.8 shows the relationship between unemployment and inflation in Australia from 1978 to 2015. It also shows the relationship between the underemployment estimates provided by the Australian Bureau of Statistics and the annual inflation rate for the same period. The equations shown are the simple regressions, depicted graphically by the solid lines. The graph suggests the negative relationship between inflation and underemployment is stronger than the relationship between inflation and unemployment. More detailed econometric analysis confirms this to be the case.

The inclusion of underemployment in the Phillips curve specification helps explain why low rates of unemployment were not inflationary in the period leading up to the GFC. It suggests that changes in the way the labour market operates, with more casual work and underemployment, have been significant in explaining the impact of the labour market on wage inflation and general price level inflation.

Figure 18.8 The inflation rate, unemployment and underemployment, Australia, quarterly data, 1978–2015



Source: Authors' own. Data from Australian Bureau of Statistics, Consumer Price Index and Labour Force.

Conclusion

In this chapter we have reviewed three theories of inflation rate determination, namely the Phillips curve trade-off, the expectations augmented Phillips curve and the hysteresis model of the Phillips curve trade-off. We have shown that the expectations augmented Phillips curve specification rejected the capacity of Keynesian aggregate demand management to achieve sustainable cuts in unemployment, under the Monetarist assumption that the coefficient on the inflation expectations variable was unity.

The empirical evidence has challenged the concept of a cyclically invariant NAIRU, which underpins the Monetarist policy prescription. Also, the empirical estimates of the NAIRU are related to the lagged unemployment rate rather than any structural features of the economy, which are alleged to impact on the sustainable rate of unemployment (NAIRU). On the other hand, the hysteretic model of inflation has shown that total unemployment is an inadequate proxy for excess demand in an inflation model. Researchers need to differentiate between short-term and long-term unemployment and also incorporate underemployment into their excess demand proxy. The changing composition of unemployment over the business cycle explains why the macroequilibrium rate of unemployment (MRU) is cyclically sensitive, which gives rise to the phenomenon known as hysteresis. Both the Phillips curve trade-off and the hysteretic model are consistent with stimulatory policy leading to lower unemployment, albeit with a higher rate of inflation.

References

Brown, A.J. (1955) *The Great Inflation 1939–1951*, Oxford: Oxford University Press.

Friedman, M. (1968) "The Role of Monetary Policy", *The American Economic Review*, 58(10), March, 1–17.

Mitchell, W.F. (1987) "NAIRU, Structural Imbalance and the Macroequilibrium Unemployment Rate", *Australian Economic Papers*, 26(48), June, 101–118.

Phelps, E.S. (1968) "Money-Wage Dynamics and Labor-Market Equilibrium", *Journal of Political Economy*, 76: 678–711.

Phillips, A.W. (1958) "The Relation between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957", *Economica*, 25(100), 283–299.

Samuelson, P.A. and Solow, R.M. (1960) "Analytical Aspects of Anti-Inflation Policy", *American Economic Review*, 50(2), Papers and Proceedings of the Seventy Second Annual Meeting of the American Economic Association, May, 177–194.

Thurow, L.C. (1983) *Dangerous Currents: The State of Economics*, New York: Random House.

Endnotes

1. This is called 'Ricardian equivalence', whereby any increase of government spending is offset by a reduction of private spending because saving must rise to pay the taxes that will be imposed today or in the future to 'pay for' that government spending.
2. *Homo economicus* (economic man) usually features in orthodox microeconomic theories in which agents are depicted as both rational and self-interested. They usually make optimal decisions in the pursuit of their subjectively defined ends.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

19 FULL EMPLOYMENT POLICY

Chapter Outline

Learning Objectives

- Recognise that full employment should be the primary macroeconomic policy goal in a society that pursues internationally recognised human rights with social justice.
- Assess other policies that have been adopted to promote employment.
- Acknowledge that current policy settings in most developed economies are based on the inflation rate as the primary policy target which is addressed by an unemployed buffer stock.
- Analyse the economic consequences of implementing an employment buffer stock (Job Guarantee).

19.1 Introduction

In Chapter 17, we discussed how distributional conflict between the claimants of real income could trigger inflation if the competing nominal claims (wages, profits) exceeded the actual amount of nominal income produced in each period. We saw how this conflict could be triggered by rising wage aspirations from workers, rising profit rate aspirations from price setters (firms), and exogenous squeezes on available national income, arising from, for example, an imported raw material price rise.

The underlying dynamics of the capitalist system are driven by the target rates of profit determined by firms. In this context, workers' behaviour may lead to higher unemployment if they successfully achieve wages so high that the capacity of firms to achieve the target rate of profit is undermined. This would occur because of a reduction in effective demand if firms reduced investment spending in response to the squeeze on the rate of profit.

Alternatively, firms could raise prices to protect profits and that may lead to a price wage inflation spiral. An inflationary spiral originating from either demand pull or cost push forces requires certain aggregate demand conditions to be maintained if that spiral is to continue. As we saw in Chapter 17, this observation means that the concept of supply side inflation blurs into the concept of demand pull inflation, although their originating forces might be quite different.

The main focus of this chapter is the analysis of two approaches which are designed to achieve sustained low and stable inflation (inflation proofing). We construct the discussion in terms of a comparison between two types of **buffer stocks**, both of which are created by government policy aimed at avoiding aggregate demand pressures that might fuel an inflationary spiral.

The **two buffer stocks** that we will compare are:

- **Unemployment buffer stocks**: Under a natural rate of unemployment (NRU) (also referred to as a non-accelerating inflation rate of unemployment (NAIRU)) regime, inflation is controlled using tight monetary and fiscal policy, which leads to a buffer stock of unemployment. This is a very costly and unreliable mechanism for policymakers who are trying to achieve price stability.
- **Employment buffer stocks**: The national government exploits the fiscal power embodied in a fiat currency-issuing system to introduce full employment based on an employment buffer stock. The Job Guarantee (JG) model is an example of this type of policy approach.

Both of these buffer stock approaches to inflation control introduce so-called **inflation anchors**. In the NAIRU case, the anchor is unemployment, which serves to discipline the labour market and prevent inflationary wage demands from being pursued. Under a JG, the inflation anchor is provided in the form of an unconditional, fixed wage employment guarantee provided by the government.

In this chapter, we first elaborate on why full employment should be the key macroeconomic policy goal. Second, we briefly examine the alternative approaches that have been taken in the past to deal with chronic unemployment. Finally, we outline and contrast the two buffer stock schemes which are designed to control inflation. We show that only a JG approach provides an employed buffer stock that promotes both full employment and price stability.

19.2 Full Employment as the Policy Goal

In our discussion of public purpose in Chapter 1, we noted that in a modern capitalist economy, access to employment is required for full participation in society. Employment, especially in formal sector jobs, not only integrates individuals into networks linked to the workplace, but also into the social and political environment more generally.

On the other hand, it has been well documented that sustained unemployment imposes significant economic, personal and social costs that include:

- Loss of current national output and income.
- Social exclusion and the loss of freedom.
- Skill loss.
- Psychological harm.
- Ill health and reduced life expectancy.
- Loss of motivation.
- The undermining of human relations and family life.
- Racial and gender inequality.
- Loss of social values and responsibility.

Thus, a macroeconomic policy that uses unemployment to promote macro stability not only forces those who are already disadvantaged to bear most of the costs, but also undermines social cohesion. Joblessness is usually concentrated among groups that suffer other disadvantages: racial and ethnic minorities, immigrants, younger and older individuals, women (especially female heads of households with children), people with disabilities, and those with lower educational attainment. Lack of employment is strongly correlated with poverty and social isolation.

The **United Nations Universal Declaration of Human Rights** includes the right to work, not only because it is important in its own right, but also because many of the other economic and social entitlements proclaimed to be human rights cannot be secured without paying jobs. Also **full, productive and decent employment** is one of the UN's **Millennium Development goals.** Amartya Sen (1997) supports the right to work because the economic and social costs of unemployment are staggering, with far-reaching consequences beyond the single dimension of a loss of income (see Chapter 1 for a more detailed discussion of these issues).

Markets are not necessarily good at securing the economic and social entitlements proclaimed to be human rights in the Universal Declaration. This is why extra-market policy – policy beyond that relating strictly to, or driven by, the marketplace – has been needed to safeguard a variety of human rights. Within neoclassical theory, unemployment and poverty are generally assumed to be the necessary costs of maintaining macroeconomic stability, especially price and exchange rate stability. This raises important questions. Should a nation fight inflation by keeping a portion of its population unemployed and impoverished? Are there other tools available to achieve these ends? In particular, should policymakers accept some inflation and currency depreciation in order to eliminate unemployment and poverty? Hence, at the very least, safeguards are required to protect the minority which suffers large concentrated costs in the form of unemployment as a consequence of a policy that leads to the benefits of lower inflation accruing to society as a whole.

There are strong ethical arguments against using poverty and unemployment as the primary policy tools to achieve price and exchange rate stability, especially given that as mentioned earlier the burden of poverty and unemployment is not shared equally. And even if price and currency stability are highly desired and beneficial for all citizens, it is doubtful that a case can be made for their status as a human right on a par with the right to work.

Unemployment as a stabilisation tool not only violates various human rights, including the right to employment, it also sacrifices economic performance by generating redundant human resources. Indeed, with some notable exceptions, those countries with the highest rates of underutilised labour resources tend to be nations with the highest poverty rates.

Only government can guarantee the right to a job because markets have not, and cannot, operate at anything approaching true, full employment on a consistent basis without direct job creation on a large scale. Only the government can create an infinitely elastic demand for labour by offering to hire all who cannot otherwise find employment because it does not need to take account of narrow market efficiency concerns due to its capacity to issue and spend the sovereign currency. Private firms only hire the quantity of labour needed to produce the level of output that is expected to be sold at a profitable price. Government can take a broader view by promoting the public interest, including the right to work. For these reasons, government must play a role in providing jobs to achieve social justice. A JG programme can secure the right to work, but with minimal undesired impacts on wages, prices, government fiscal policy, and the value of the currency.

Forstater (2006) has argued that it is difficult to conceive of a policy goal that secures a greater range of social and economic rights than the achievement of full employment. His "fundamental welfare theorem of political economics" says "there is no single policy that carries with it more potential benefits than true full employment, or a guaranteed job for everyone ready and willing to work" (Forstater, 2006: Slide 2).

In addition to income, employment also yields useful production and recognition for doing something worthwhile. While economists usually focus only on the economic multiplier, there are also **social multipliers** associated with job creation. These benefits include decreased crime and drug use; enhanced family and community cohesion; improved economic security, education, and healthcare; protection for the disadvantaged; environmental protection; improved local and state government budgets; more equal distribution of consumption, income, wealth, and power; induced investment in poor communities; and promotion of social and political stability.

While economic growth and development are desirable, they do not ensure either **full or decent employment**. Alternative strategies for promoting full and decent employment will be explored in the next section. Certainly it is necessary to attack problems of unemployment, underemployment, and insufficient pay using a variety of programmes and policies. These should include both private and public initiatives. However, it will be argued that the private sector will not be able to provide for full, productive, and decent employment for all, even with substantial support by government through the promotion of private sector job creation. Only the introduction of a safety net as provided by a JG can protect the human right to employment.

19.3 Policies for the Promotion of Employment

Behaviouralist, structuralist, and Keynesian approaches

A range of strategies have been adopted to address the problem of joblessness, of which the most important are behaviouralist (problems with the individuals who are unemployed), structuralist (for example, skills mismatch), and job shortage.

Kaboub (2008) provides an historical overview of attitudes of economists toward the unemployed, from Petty (1662: 160) (the unemployed "ought neither to be starved, nor hanged, nor given away" because they represent a resource that could be used in public employment to enrich the nation) to Beveridge (1945:10) who wanted full employment, defined as "having always more vacant jobs than unemployed", and to lesser-known advocates of JG-type schemes such as Pierson (1980) and Wernette (1945). He also surveys full employment strategies that have been adopted, such as the New Deal in the Great Depression-era USA (1933–36), the Swedish post-war model (1950–75), job creation schemes in India (2006–), Argentina (2002–), and more recently the French pilot programme to create jobs for laid-off workers.

Public attitudes and policy have generally emphasised behavioural and structural problems. This leads to policies that try to motivate and train the unemployed, together with the promotion of greater flexibility (such as wage flexibility) that would reduce labour market frictions. Here unemployment is conceptualised as largely an individual problem arising from behavioural deficiencies, rather than being a systemic macroeconomic problem.

However, if the problem is a job shortage, all that these policies can do is to redistribute unemployment within the unfortunate group who are blamed for their joblessness. In an expansion, some of those job seekers who are not hired do have the characteristics identified with the behaviouralist and structuralist arguments (since employers hire those with the most desirable characteristics first), hence concealing the true problem: a chronic job shortage.

While Hyman Minsky is best known for his work on financial fragility, he also argued that any public policy that favours education and training over job creation puts the cart before the horse and is unlikely to succeed (Minsky (1986, for example)).

First, it lays the blame on the unemployed, which can be demoralising and seems to validate existing public perceptions regarding the undesirable characteristics that are supposedly endemic within the disadvantaged population. The message is that the poor must change their characteristics, including their behaviour, before they deserve to work. However, those without jobs might not view such changes as desirable or even possible.

Second, it can require a long time to see results; the gestational period to produce a worker is at least 16 years for developing nations and 25 years or more for highly developed nations. Further, as structuralists recognised, a dynamic economy is always leaving old skills behind and demanding new ones. At any point, there will be a permanent, sizeable, pool of those with inappropriate skills and education, even if many individuals are able to transition out of the pool in a timely fashion.

Third, as mentioned, there is the danger that the retrained will face a job shortage so that at best they simply displace previously employed workers who will join the ranks of the unemployed.

For these reasons, jobs that can take workers as they are, regardless of their skills, education, or personal characteristics, must be made available. The upgrading of the workers' characteristics would be the second step, with most of the necessary training occurring on the job. The unemployed need jobs, not merely the promise of a job for those who successfully reform themselves.

Note also that if welfare (including unemployment compensation) is offered as a substitute for a job, this has negative impacts on self-esteem, on public perceptions of the unemployed, and on human capital (skills and experience) that deteriorates through lack of use. For these reasons, providing welfare rather than work to those who want to work is not only an admission of defeat (accepting that the labour market fails to provide enough jobs), but also wastes resources and generates social costs.

After the Second World War, the notion that **Keynesian policies** could keep aggregate demand at a sufficiently high level to promote robust growth came to dominate Western thought. Further, it was believed that high growth would keep unemployment low and thereby reduce poverty rates. Aggregate demand was sustained in the West (especially the US) through spending on defence and public infrastructure investment, and through the favourable treatment of private investment. The high growth strategy was supplemented by a combination of behaviouralist and structuralist labour market policies plus welfare (itself a labour market policy, in the sense that one of the goals of welfare for families with children and retirement income for the aged was to reduce the size of the labour force). This was deemed necessary because growth was leaving behind pockets of hardship among disadvantaged groups, with poverty becoming regionally and racially concentrated.

During the early post-war period, economic growth was maintained at an above-average rate, and unemployment and poverty rates seemed to fall; apparently validating the Keynesian approach.

However, these policies would not eliminate unemployment and poverty because they failed to prioritise job creation. At best, they redistributed joblessness. Further, a high growth strategy actually favoured the more advanced sectors of the economy-that is, those with highly skilled and well-paid workers-which increased income inequality. Finally, policy that favours high investment would prove to be unsustainable because it generated macroeconomic instability, evidenced by inflation, currency devaluation, and financial fragility.

For these reasons, policy-induced recessions would be required to try to restore conditions favourable to macro stability. This led to a **go/stop/go pattern** of using fiscal stimulus in downturns and tight fiscal policy near a business cycle peak. As a result, expansions were curtailed long before a sufficient supply of jobs was created in order to achieve full employment and address poverty.

Keynesian policy fell out of favour during the **stagflationary** 1970s. In March 1973, the major currencies floated after the collapse of the fixed exchange rate (Bretton Woods) system. The social safety nets adopted in the early post-war period had been either dropped or underfunded, and neoliberalism (called neoconservatism in the US) played a growing role in the developed economies, such as the US, UK and Australia.

Finally, recessions and financial crises returned after 1970 when the belief that high growth and low unemployment were inconsistent with price stability came to dominate policy formation. Financial fragility appears to have risen over time, as evidenced by increasingly frequent and severe domestic and international financial crises.

Policymakers have turned away from the use of fiscal policy to promote growth, and have largely relied on monetary policy. Monetary policymakers in turn generally deny responsibility for maintaining high employment and growth, except to the indirect extent that low inflation promotes a strong economy. This prioritisation of low inflation over the achievement of high levels of employment represents the adoption of the **unemployment buffer stock** approach to policy (see Section 19.4).

Substantial controversy surrounds all of these issues, but it is commonly accepted that attempts to fine-tune the economy through Keynesian aggregate demand manipulation have proven to be largely unsuccessful. Even if these policies had been successful, there is little political will to return to them today. However, the chosen replacement – neoliberalism – has not succeeded either.

There are a further two policies for the promotion of employment: indirect job creation through incentives given to the private sector, and direct employment by government. Both strategies have been tried in many economies.

Private sector incentives

There are several drawbacks to subsidised employment in the private sector. First, government needs to ensure that firms use the subsidies to create jobs, rather than to reduce the private costs of existing employment. In a dynamic economy with jobs continually created and destroyed, this is difficult to police because profit-seeking firms would want to use government funds to subsidise existing jobs, resulting in public spending without the creation of any net new jobs.

Second, as unemployment is concentrated among disadvantaged workers, the policy should encourage firms to employ individuals they would not otherwise have hired. Again, this is difficult to monitor because firms would want to hire those job seekers with the most desirable characteristics that are allowed under programme rules, rather than those with average (or lower) characteristics. Further, there is the danger that firms will hire eligible workers, displacing workers with similar characteristics who are considered marginal in some way or another.

Third, there are questions about the time span permitted for eligibility. One goal of the programme should be to take workers with little experience or skills and prepare them for unsubsidised employment. However, if workers are permitted to stay in the programme for only a specified period, there is a strong incentive for employers to replace these workers at the end of their period of eligibility with newly eligible and subsidised workers, rather than retaining the group who were initially hired under the job creation scheme but who no longer bring in a subsidy. Workers who are forced to leave the programme might not find unsubsidised work.

Fourth, the setting of the wage subsidy is not necessarily simple. The subsidy required to induce firms to hire a new worker presumably varies according to the perceived shortfall of the worker's employability relative to the pool of workers from which the firm normally recruits. A sliding scale subsidy might be most effective, but it could be difficult to establish the proper subsidy rates. The required subsidy will also vary according to the firm's need for new workers. In an economic boom, a small subsidy might be sufficient to induce an employer to hire one more worker than the firm would have otherwise employed. In a deep recession, even a 100 per cent wage subsidy might not induce a firm to hire one more worker.

Finally, the payment of wage subsidies necessarily leads to some distortion of the market. Some firms will be able to take advantage of the scheme, while others will not. Some existing employees will have to compete with subsidised labour while others will not. Some lines of production will increase output because of additional workers while others will not, and so on.

While none of these potential problems, even if taken together, necessarily implies that a programme of wage subsidies should not be tried, the potential problems would seem to lead to the conclusion that such a programme probably cannot by itself solve the problem of joblessness and ensure the right to work.

Of course, private sector subsidies will not work without a sufficiently developed private sector which is capable of offering employment to a significant portion of the population. In some developing nations, especially in rural regions, such a policy will have limited potential application. Only direct job creation at a living wage by government will provide a sufficient supply of non-agricultural work to reduce joblessness.

Direct job creation by government

We conclude that raising aggregate demand, increasing human capital, and raising the incentives to private employers will fall short of ensuring the right to work. While each of these policies might be useful in its own right, they must be supplemented by direct job creation by government. Most governments engage in some form of job creation for the purpose of relieving unemployment. Arguably, the nations that achieved anything close to full employment in the post-war years used a variety of programmes to keep unemployment low. They all maintained, in one form or another, a buffer of jobs that were available to the least skilled workers, who otherwise were likely to be unemployed.

The main criticisms of government job creation schemes are that, unlike the JG, they typically do not provide ongoing employment under normal working conditions, and that their coverage is limited to particular groups, such as rural workers (Mahatma Gandhi National Rural Employment Guarantee Scheme in India), heads of household (Jefes de Hogar programme in Argentina) and youth (Youth Guarantee in EU countries).

In the next two sections we examine the two buffer stock regimes that are designed to promote price stability. The first of these is not consistent with the pursuit of full employment. Indeed, by design it uses unemployment to stabilise prices. The second stabilises prices as it pursues full employment.

As we demonstrate, a **Job Guarantee** (JG) is at the centrepiece of MMT reasoning. It is neither an emergency policy nor a substitute for private employment, but rather would become a permanent complement to private sector employment. A direct job creation programme can provide employment at a basic wage for those who cannot otherwise find work. No other programme can guarantee access to jobs at decent wages. Further, the JG approach has the advantage that it simultaneously deals with the main objection to full employment: the argument that the maintenance of full employment causes unsustainable rates of inflation.

19.4 Unemployment Buffer Stocks and Price Stability

There have been two striking developments in macroeconomics over the last forty years or so. First, a major theoretical revolution occurred whereby Keynesianism was supplanted by Monetarism and subsequently by its more modern variants. Second, the demise of Keynesianism in the early 1970s was accompanied by persistently higher unemployment rates, and during the GFC they rose even higher. Further, mean economic growth rates have fallen during this neoliberal period.

As we saw in Chapters 11 and 12, prior to Keynes' *General Theory*, unemployment at the aggregate level was seen by many orthodox economists as a temporary deviation from equilibrium which was due to labour market frictions or other market disruptions. Keynes changed the discourse to one that blamed aggregate unemployment on insufficient aggregate demand. This led in the post-war period to the belief that Keynesian demand management policy was the proper response. Unemployment rates were usually below two per cent throughout this period in many countries (although somewhat higher in the USA, especially among minority groups).

However, in the early 1970s, the Phillips curve trade-off appeared to break down with countries experiencing stagflation. In time, most of the mainstream economists rejected demand management and returned to the older pre-Keynesian belief that some level of aggregate unemployment is: (a) temporary and due to shocks; (b) optimal because it is voluntary; and/or (c) the necessary cost of promoting stability. Thus the concept of full employment as a genuine policy goal was abandoned with the introduction of the **natural rate of unemployment hypothesis**, which has become a central plank of current mainstream thinking.

This hypothesis asserts that only this 'natural' unemployment rate is consistent with stable inflation. Therefore, there is no discretionary role for aggregate demand management and only microeconomic changes can reduce the natural rate of unemployment. Thus, the policy debate increasingly concentrated on deregulation, privatisation, and the reduction of welfare state provisions within an environment of tight monetary and fiscal regimes.

Under inflation-targeting (or inflation-first) monetary regimes, central banks have shifted their policy emphasis. They now conduct monetary policy to meet an inflation target,[1] and arguably have abandoned any obligation to support a policy environment which achieves and maintains full employment. This almost exclusive focus on maintaining price stability on the back of an overwhelming faith in the NAIRU ideology has marked the final stages of the abandonment of earlier full employment policies.

Since the mid-1970s unemployment in many economies has remained at high levels. In addition, low quality, casualised work has emerged in the face of the persistently deficient demand for labour hours by employers. Thus, underemployment has replaced some unemployment.

As we saw in Chapter 17, underemployment acts in a similar way to unemployment by operating as a disciplining force on workers' wage aspirations and demands. It weakens the capacity of workers to secure nominal wages growth. Thus, short-term unemployment and underemployment temporarily balance the conflicting demands of labour and capital by disciplining the aspirations of labour so that they are compatible with the profitability requirements of capital. Similarly, low market demand, the analogue of high unemployment when workers' incomes fall, suppresses the ability of firms to increase prices to protect real margins.

Thus by inducing labour slack into the economy, inflation targeting supported by passive fiscal policy leaning towards austerity has created what Karl Marx called a **reserve army of the unemployed** and this has reduced the chances of an inflationary spiral emerging from the wage bargaining process.

The NAIRU approach to price stabilisation is therefore based on government spending being subject to a **quantity rule**. This means that the government plans for a quantity of the local currency to be spent at prevailing market prices to prosecute its socio-economic programme. Spending overruns are usually met with cutbacks in an attempt to meet the fiscal targets.

We have seen significant shifts in the distribution of national income towards profits since the mid-1980s as real wage growth has lagged behind productivity growth. This redistribution of national income has overridden the previous outcomes when strong trade unions met on more equal terms with employer groups to determine a distribution of national income that would be acceptable to both sides of the bargaining process.

But with trade unions weaker as a result of shifting industry composition towards services, smaller public sectors and a rise in anti-union legislation, slower wages growth and low inflation in recent years have raised fears of deflation in developed economies. As a consequence, the use of unemployment as a tool to suppress price pressures has, based on the OECD experience since the 1990s, been successful as measured by inflation outcomes.

The empirical evidence is also clear that most OECD economies have not provided enough jobs since the mid-1970s and the conduct of monetary policy has contributed to this malaise. Central banks around the world have forced the unemployed to engage in an involuntary fight against inflation and the fiscal authorities in many cases have further worsened the situation with complementary austerity measures.

These costs are very large and have long-term consequences. In terms of the goals of macroeconomic policy, they also present a major conflict. As we have learned, a central idea in economics whether it be microeconomics or macroeconomics is efficiency: that is, getting the best out of what is available. We have discussed the difficulties that economists have in defining such a concept, and its ideological dimensions.

But economists could put aside their differences and agree that at the macroeconomic level, the efficiency frontier should be defined in terms of full employment. The major debate, which we covered in Chapter 17, concerned how we might define full employment. It is a fact that full employment should be a central focus of macroeconomic theory and policy.

Certainly mass unemployment, with hundreds of thousands or even millions of workers not producing any output or national income, violates any reasonable definition of macroeconomic efficiency.

Further, persistently high unemployment not only undermines the current welfare of those affected and slows down the growth rate in the economy below its potential. It also reduces the medium- to longer-term capacity of the economy because the accompanying erosion of skills and lack of investment in new capacity means that future productivity growth is likely to be lower than if the economy was maintained at higher rates of activity.

The NAIRU approach to price stabilisation is typically complemented by policies which impose obligations on the unemployed in exchange for their welfare benefits. The unemployed usually have to demonstrate that they are engaged in activities which are alleged to increase their likelihood of securing employment. These include participation in training programmes and applying for a minimum number of jobs per month. Again, at best, in a job-constrained economy these policies will merely redistribute unemployment among this disadvantaged group.

The key question to ask advocates of the unemployment buffer stock approach to inflation control is whether the economy, once deflated by restrictive aggregate demand management, can be safely restarted. If the underlying causes of the inflation are not addressed, demand expansion will merely reignite tensions over the distribution of income between wage earners and profit recipients and a wage-price outbreak is likely.

> **As a basis for policy, the NAIRU approach has major limitations, because it addresses the symptoms and not the causes of inflation, and as a consequence, provides no firm basis for sustained full employment and price stability.**

In short, its success as an inflation anchor requires a pool of persistently high unemployment.

The disciplining power of unemployment requires that the unemployed continue to constitute a threat to those still in work so that the employed will moderate their wage demands. However, over time, the threat from the unemployment pool starts to wane as the unemployed endure skill losses, while suffering lengthening periods of unemployment, and firms introduce new technologies and processes. This is referred to as hysteresis (see Chapter 18). In this case, the so-called NAIRU has to be pushed higher and higher by contractionary fiscal and monetary policy for the same degree of threat to be maintained.[2]

On any reasonable grounds, this approach to price stability is very costly and ultimately unworkable in a modern economy. High and sustained levels of unemployment undermine the social and political stability of a nation, which creates unintended costs that go far beyond those that are itemised above.

Measuring the costs of unemployment buffer stocks

Under inflation-targeting monetary policy regimes, central bankers use the persistent pool of unemployed (and other forms of labour underutilisation such as underemployment) as a buffer stock to achieve a desirable inflation outcome. If their inflation outlook rises above their target rate, they will induce higher rates of unemployment by increases in interest rates until they are satisfied that their inflation target is being met (see Chapter 23).

While some extreme free market economists, who still consider rational expectations to be a reasonable assumption, will deny any real output effects, most economists acknowledge that any disinflation engendered by this approach will be accompanied by a period of reduced output and increased unemployment (and all its related social costs) because a period of (temporary) slack in the economy is required to break inflationary expectations.

The real question then is how large are the output losses following a discretionary disinflation? There is overwhelming evidence to suggest that the cumulative costs of this strategy in real terms have been substantial.

Research has been undertaken to calculate the **sacrifice ratio** associated with the implementation of the unemployment buffer stock policy. This is defined as the ratio of the accumulated loss of output during a disinflation episode expressed as a percentage of initial output divided by the overall reduction in the inflation rate. For example, if the sacrifice ratio were two, it would mean that a one-point reduction in the trend inflation rate is associated with a cumulative GDP loss equivalent to two per cent of initial output.

Figure 19.1 is a simple graphical depiction of the sacrifice ratio concept. The total output loss resulting from actual output falling below potential output is depicted by the shaded area. We have deliberately shown output as resuming at its potential (long-run) level at the end of the disinflation period (defined as the period between the peak inflation and the trough inflation). This is the normal assumption adopted in empirical studies.

On the other hand, the concept of **persistence** means that actual output remains below its potential level after the disinflation period has finished. The longer this output gap exists, the longer is the persistence.

In this context, **hysteresis** refers to the permanent losses of potential output that arise as a consequence of the disinflation policy. Thus the growth of potential output is permanently lowered due to the collateral damage of low confidence among firms which curtails investment. This in turn limits the long-term growth of actual output.

The important point is that, in order to accurately estimate the sacrifice ratio, researchers must not only consider the short-term losses, but also the longer-term losses arising from persistence and hysteresis.

Figure 19.2 stylises the impacts of persistence and hysteresis arising from the adoption of a disinflationary policy. From the inflation peak, real output falls immediately as before. However after a time, the reduced levels of economic activity erode confidence among consumers and firms. Consumers fearing even higher unemployment restrict consumption spending and firms respond to the lack of sales by cutting investment plans.

Two impacts occur as a result: (a) the potential real output path falls (after Trough + x quarters on Figure 19.2), reducing the growth capacity of the economy; and (b) actual real output deviates from its potential path for much longer than otherwise would have been the case.

The real output losses are much greater than those shown in Figure 19.1. Eventually actual and potential output paths may converge, but at that point there is less output and national income, and almost certainly persistently higher unemployment.

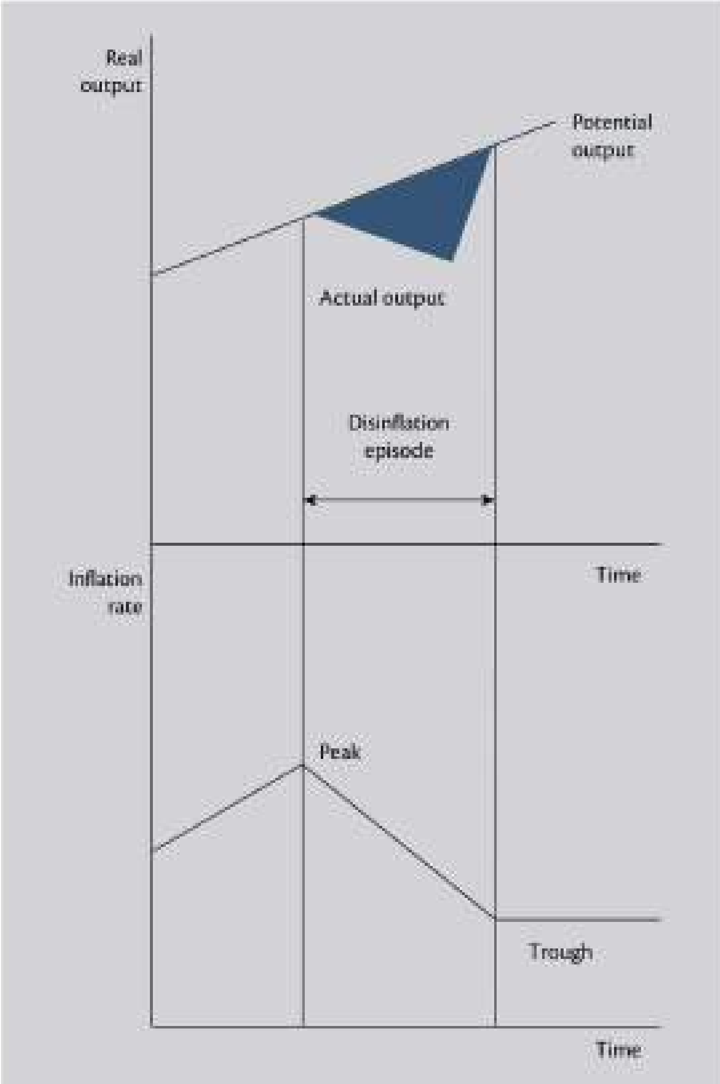
Mass unemployment was initially caused by the deliberate cutting of aggregate demand due to the contractionary policy stance of the government, but the subsequent expansion of output can become capacity-constrained as a result of a slow growing or even falling potential output level due to the weak inducement to invest.

The fiscal austerity policies pursued by governments during the Global Financial Crisis also had this impact. It would then be much harder to restore robust growth because it takes longer to ensure there is also potential capacity to support it without triggering inflation.

Mitchell and Muysken (2008) drew on an extensive literature analysis and their own empirical work to conclude the following:

- Formal econometric analysis does not support the case that inflation targeting delivers superior economic outcomes in terms of reducing the costs of disinflation. Both targeters and non-targeters enjoy variable outcomes, and there is no credible evidence that inflation targeting improves performance as measured by the behaviour of inflation, output, or interest rates.

Figure 19.1 The sacrifice ratio and disinflation episode



- There is no credible evidence that central bank independence and the alleged credibility bonus that this brings lead to faster adjustment of inflationary expectations to the policy announcements. There is no evidence that targeting affects inflation behaviour differently.
- Estimates of sacrifice ratios have confirmed that disinflations are not costless. Significantly, the average estimated GDP sacrifice ratios increased over time, from 0.6 in the 1970s to 1.9 in the 1980s and to 3.4 in the 1990s. That is, on average a reduction of trend inflation by one percentage point resulted in a 3.4 per cent cumulative loss in real GDP in the 1990s.
- Australia, Canada, and the UK, which announced formal policies of inflation targeting in the 1990s, do not have substantially lower sacrifice ratios compared to G7 countries that did not announce such policies. Australia does appear to have recorded a lower average ratio during the targeting period than in the 1980s. However, this figure is not lower than the average for all previous periods. Canada recorded a higher sacrifice ratio in the 1990s of 3.6. The ratio for the UK during inflation targeting was significantly higher at 2.5 (relative to quite low sacrifice ratios in previous periods). Italy, Germany, Japan and the US, averaged 0.6, 2.3, 2.9 and 5.8, respectively.

The evidence is clear that inflation-targeting countries have failed to achieve superior macroeconomic outcomes in terms of output growth, inflation variability and output variability. Moreover, there is no evidence that inflation targeting has reduced the persistence of inflation.

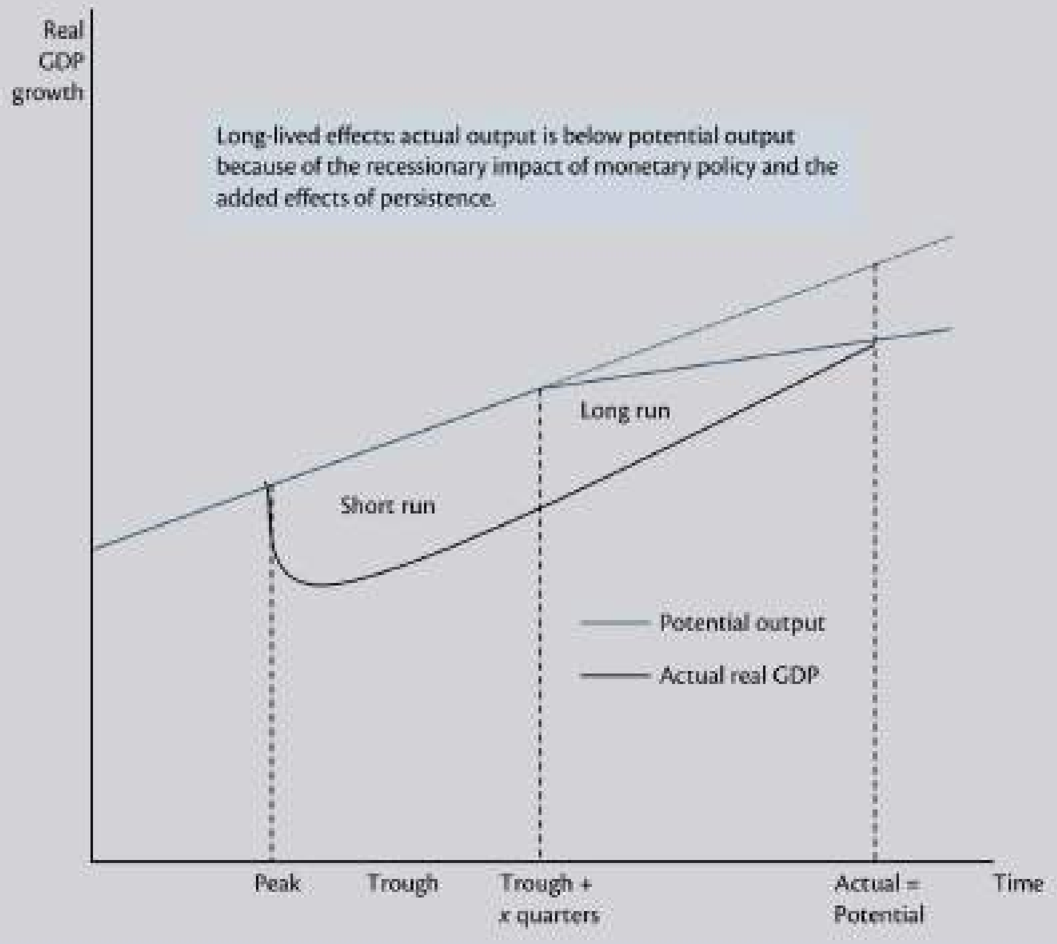
Other factors have been more important than targeting *per se* in reducing inflation. Most governments adopted fiscal austerity measures in the 1990s in the mistaken belief that fiscal surpluses are the exemplar of prudent economic management and provide a supportive environment for monetary policy. The fiscal cutbacks had adverse consequences for unemployment and created additional slack in the labour market. Labour underutilisation (defined more broadly to include among other things, underemployment) rose in these countries.

Further, other factors (the comprehensive shift to active labour market programmes, welfare-to-work reform, dismantling of unions and privatisation of public enterprises) also helped to keep wage pressures down.

It is clear from statements made by various central bankers (in addition to their formal obligations) that a belief in the absence of a long-run trade-off between inflation and employment embodied in the NAIRU literature has led them to pursue an inflation-first strategy at the expense of unemployment.

Disinflations are not costless, irrespective of whether targeting is used or not. An average sacrifice ratio of 3.5 in the 1990s meant that any attempt to bring down inflation by one percentage point resulted in a cumulative loss in GDP of 3.5 per cent on average. In terms of unemployment, the latter can be interpreted roughly speaking as a cumulative increase by seven percentage points.

Figure 19.2 Sacrifice ratios with persistence and hysteresis



The increase in the sacrifice ratio over time means that the Phillips curve has become flatter with a greater increase in unemployment required for a given cut in the inflation rate. Thus a consequence of inflation targeting is that the costs of disinflation have increased.

Franco Modigliani (2000: 3), one of the economists who coined the term NAIRU, reflected on the legacy he had created:

> *Unemployment is primarily due to lack of aggregate demand. This is mainly the outcome of erroneous macroeconomic policies … [the decisions of Central Banks,] inspired by an obsessive fear of inflation, … coupled with a benign neglect for unemployment … have resulted in systematically over-tight monetary policy decisions, apparently based on an objectionable use of the so-called NAIRU approach. The contractive effects of these policies have been reinforced by common, very tight fiscal policies.*

One of the major problems of inflation targeting as a policy paradigm is that it has been accompanied by a view that fiscal policy must be passive and not compromise the inflation target. As a result, economies have tolerated persistently high rates of labour underutilisation despite having achieved low inflation.

As noted earlier in the chapter, persistent unemployment results not only in massive losses of real output and national income, but other real costs are also endured by the nation, including the depreciation of human capital, family breakdown, increasing crime, and increasing medical costs.

These additional costs, in particular the depreciation of human capital, also mean that the effectiveness of the unemployed pool as a price anchor deteriorates over time, with ever-larger numbers of fresh unemployed or underemployed being required to function as a price anchor that stabilises wages.

Given the scale of these costs, it is unlikely that using a persistent pool of unemployed or casualised underemployed is the most effective way to achieve price stability.

19.5 Employment Buffer Stocks and Price Stability

Given the importance placed on access to employment for those of working age who want a job, a better alternative to an unemployment buffer stock to achieve price stability would be to utilise an employment buffer stock, as long as price stability is not compromised.

In this section, we outline an employment programme for these unemployed persons as an activity floor in the real sector, which both anchors the general price level to the wage of this (currently unemployed) buffer stock of employed labour but also can produce useful output with positive supply side effects.

REMINDER BOX

Recall that the MMT approach argues that the imposition of taxes by the currency-issuing government generates a demand for the currency. The currency's value is determined by what must be done to obtain it. The currency will be worth the amount of labour it can buy on the margin, which is the wage paid in the JG programme. The wage and benefit package in the JG programme sets a standard for what must be done to earn currency-denominated income.

To be sure, the anchor is not tight. Some people earn a level of income, either from the private sector or from the government, which is disproportionate to their level of endeavour. If everyone could get income while doing nothing (if money grew on trees!) then the currency's value would approach zero. However, in the real world, the government's currency does not grow on trees and most people have to do something to get it. For that reason, at the margin, currency is valuable.

Between 1945 and the mid-1970s, Western governments realised that with deficit spending supplementing private demand, they could ensure that all workers who wanted to work could find jobs. Although private employment growth was relatively strong during this period, governments were important employers in their own right, and also maintained a buffer of jobs for the least skilled workers. For example, jobs were available in the major utilities, the railways, local public services, the army, and major infrastructure functions of government. By absorbing workers who lost jobs when private investment declined, governments acted as an economic safety valve.

British economist Paul Ormerod (1994: 203) noted that the economies that avoided high unemployment in the 1970s maintained a "sector of the economy which effectively functions as an employer of last resort, which absorbs the shocks which occur from time to time, and more generally makes employment available to the less skilled, the less qualified." He concluded that societies with a high degree of social cohesion (such as Austria, Japan and Norway) were more willing to ensure that everyone had access to paid employment opportunities.

The **employment buffer stock** approach, which is more usually referred to in the literature as the **Job Guarantee** (JG), defines a policy framework in which the government operates a buffer stock of jobs to absorb workers who are unable to find employment in the private sector.

Analogous to the central bank's function of lender of last resort, the JG functions as a buffer which employs all job seekers who have not obtained regular public or private sector jobs at a socially acceptable minimum wage. In this sense, the government acts as an employer of last resort. The jobs are available on demand.

While it is easy to characterise the JG as purely a public sector job creation strategy, it is important to appreciate that it is actually a macroeconomic policy framework designed to deliver full employment and price stability based on the principle of buffer stocks.

Under a JG, the government provides an unconditional, open-ended job offer at a given wage to anyone who desires to work. Instead of a person becoming unemployed when aggregate demand falls below the level required to maintain full employment, that person would enter the JG workforce.

This approach represents a shift from spending based on a **quantity rule** under the NAIRU approach to spending being underpinned by a **price rule**. Market forces determine the total quantity of government spending that would be required to satisfy the demand for public sector jobs at the fixed wage under a JG.

Thus the JG pool expands(declines) when private sector activity declines (expands). Hence the JG fulfils an absorption function that minimises the costs associated with the flux of economic activity when aggregate demand fluctuates. If aggregate demand declines, total demand for non-JG workers declines according to the employment requirements function we saw in Chapter 16. The workers displaced from their jobs would have a choice: accept a JG position or wait for conditions to improve in the non-JG economy.

The decisions made by these workers would be influenced by several factors. First, the government may offer workers the choice between the JG wage and unemployment benefit, with the latter being lower. Second, some workers, especially those in higher skilled positions, may receive redundancy (unemployment compensation) payments and use these to support themselves through the spell of unemployment. Economists call this response 'wait unemployment'. These workers may feel that accepting a low-skill JG job would disadvantage them professionally and thus wait for circumstances to improve. (We assume here that the JG policy does not offer an unemployment benefit, and that most displaced workers will prefer a JG position over wait unemployment. These assumptions serve to simplify the analysis; relaxing them does not alter the basic dynamics of the system).

When private economic activity picks up, workers would be bid for out of the JG pool by employers and the buffer stock of jobs would contract. Government's spending on the programme would move countercyclically, helping to stabilise aggregate employment, demand, income, and production.

The JG programme also helps to stabilise aggregate wages, because no worker's wage can fall below the JG wage. For most of the working population, the wage is usually the most important source of income. Stable aggregate wages in turn help to stabilise consumption. For that reason, this targeted approach to sustaining full employment is a powerful stabilising force for aggregate demand, output, and prices. In particular, the variation of non-JG spending will be smaller under a buffer stock employment model, which means that the stock of JG employment can be kept relatively low, albeit subject to the impact of the business cycle.

The JG wage

The buffer stock employees would be paid a living wage, which would define the level of income necessary for a full-time worker to enjoy an adequate social and material existence.

The nation's workforce would always remain fully employed, with the mix between private and public sector employment fluctuating as it responds to the spending decisions of the private sector. Since the JG wage is open to everyone, it will effectively be the national minimum wage because private employers would have to meet it in order to retain workers (except in unusual circumstances).

While it is preferable to avoid disturbing the private sector wage structure when the JG is introduced, a case can be made to offer the JG wage at a level higher than the existing private minimum if it is thought that productivity is too low in the economy. This is particularly relevant in developing economies where many market-based jobs pay wages that are below the poverty line and provide no incentives for employers to invest in more productive capital, or for workers to invest in human capital. The government would supplement JG earnings with a wide range of social wage expenditures, including adequate levels of public education, health, childcare, and access to legal aid.

The minimum wage should not be determined by the private sector's capacity to pay. It should be an expression of the aspiration of the society in terms of the lowest acceptable standard of living. Any private operators who cannot afford to pay the minimum should exit the economy.

Further, the JG policy does not replace the use of fiscal policy to achieve social and economic outcomes. Typically, the JG would be accompanied by higher levels of public sector spending on public goods and infrastructure. These supplements would be in addition to the scheme but not essential for it to function effectively.

The JG as an automatic stabiliser

The JG wage thus defines the wage floor for the economy and serves as an automatic stabiliser, complementing the tax system.

Recall that automatic stabilisation refers to the components of the government fiscal outlays and receipts which rise and fall as the economic cycle fluctuates without any explicit change in government spending or tax settings. They operate to stabilise the economic cycle by providing a floor following a fall in aggregate demand during an economic downturn and a ceiling for aggregate demand as the economy grows. At full employment, the automatic stabiliser component of aggregate demand is zero. Thus, when the economy is in decline, tax revenue falls and welfare payments rise, which expands the fiscal deficit of the government automatically. The introduction of the JG would have the same countercyclical impact. When the economy is faltering, the spending associated with the JG would rise and vice-versa when times are good.

In this regard, the JG is a superior (more powerful) automatic stabiliser than a system of unemployment benefits (under the unemployment buffer stock option) because aggregate demand slumps less and therefore the positive impact on real output is greater than would be the case if the government merely paid unemployment benefits. Further the operation of the JG sustains full employment, with all the personal and societal benefits of full employment discussed earlier.

Automatic stabilisers have the desirable characteristic of providing immediate, countercyclical spending injections (or withdrawals) when private activity fluctuates. They avoid the so-called policy lags, which relate to the time delays between the government: (i) identifying that a significant shift in private demand has occurred; (ii) designing a policy response to that shift; (iii) providing appropriate legislation to support an intervention; and (iv) executing the intervention.

In some cases, these delays can result in the major part of a new policy intervention arriving too late and operating to destabilise the cycle. For example, by the time the government has designed and implemented a new discretionary spending injection, the private sector may have already resumed normal spending growth so that the impetus provided by government spending might lead to the economy to overheat. This economic destabilisation would not occur under a JG. Workers who have become unemployed following a fall in aggregate demand can readily identify themselves to the appropriate government agency and secure a JG job.

The fixed wage offer that defines the JG policy also serves to stabilise the growth rate in money wages in the economy and thus provides a nominal anchor against inflation.

By design, a JG programme is a complement to private sector employment and to other active labour market policies, fiscal policies that aim to fine-tune total spending, and welfare or other social safety nets. A universal JG programme, one which employs anyone who is ready and willing to work, is the only type of programme that can ensure that the human right of full employment is continuously met. If the programme wage is a living wage, it also helps to ensure that other human rights are met by providing sufficient income. A properly designed programme will not only produce socially useful goods and services, but it will also promote feelings of self-worth and accomplishment among programme participants. Finally, the JG generates full employment and macroeconomic stability with the least disruption to markets.

The notion of a JG has a long history, and there are many examples of such programmes through history and across the globe, although usually enacted on a small scale or temporary basis.

Inflation control and the JG

While introducing a public sector job creation capacity to the economy, the JG is better thought of as a macroeconomic policy framework designed to ensure that full employment and price stability are maintained over the private sector economic cycle.

What are the mechanics of inflation control under a JG? In Chapter 17, we examined the way in which incompatible claims over the available real income could cause wage-price pressures to escalate into an inflationary episode as the claimants (labour and capital) attempt to defend their income shares.

In an unemployment buffer stock system, unemployment is used to discipline wage demands by workers and to soften the product market so as to discourage a profit margin push by firms as a means of curbing wage-price pressures and maintaining stable inflation.

We define the **buffer employment ratio (BER)** as:

$$BER = JGE/E \tag{19.1}$$

where JGE is total employment in the Job Guarantee buffer stock and E is total employment in the economy. **The BER rises when the JG pool expands and falls when the JG pool contracts.**

The JG approach stands in contradistinction to the NAIRU approach because instead of manipulating the employment rate by creating unemployment when wage-price pressures develop, the government manipulates the BER.

When the level of private sector activity and the distributional conflict is such that wage-price pressures form as the precursor to an inflationary episode, the government manipulates fiscal and monetary policy settings (preferably fiscal policy) to reduce the level of private sector demand.

Labour is then transferred from the inflating private sector to the **fixed wage JG sector** and the BER rises. This will eventually ease the inflationary pressures arising from the wage-price conflict.

There can be no inflationary pressures arising directly from a policy where the government offers a fixed wage to any labour not wanted by other employers. The JG sees the government **buying labour off the bottom**, in the sense that employment at the minimum wage does not impose pressure on the market sector wage structure. By definition, the unemployed have no market price because there is no market demand for their services.

By not competing with the private market, the JG would avoid the inflationary tendencies of past Keynesian policies, which attempted to maintain full capacity utilisation by 'hiring off the top' (that is, making purchases at market prices and competing for resources with all other sources of spending in the economy). In practice, these policies often focused spending on the most advanced sectors employing higher-skilled (usually unionised) workers in the defence sector, for example.

Would the incumbent workers use the decreased threat of unemployment to pursue higher wage demands? That is unlikely. First, the JG lowers the cost of hiring for firms because the JG workers do not experience the dislocation of unemployment and retain most, if not all, of their general and specific skills. Second, there might be little perceived difference between unemployment and a JG job for a highly paid worker, which means that they will still be cautious in making wage demands.

The BER conditions the overall rate of wage demands. When the BER is high, real wage demands will be correspondingly lower and the capacity of firms to push profit margins up is reduced, due to weaker product demand.

So instead of a buffer stock of unemployed being used to discipline the distributional struggle, the JG policy achieves this via compositional shifts in employment through transfers in and out of the JG pool. JG policy anchors the general price level to the price of an employed labour buffer stock, and can produce useful output with positive supply side effects.

Importantly, the JG can also deal with a supply shock (such as a rise in the price of a key non-labour raw material) that generates incompatible claims on national income that ultimately cause inflation.

The NAIRU defines the unemployment buffer stock associated with stable inflation. In a JG setting, we define the **non-accelerating inflation buffer employment ratio (NAIBER)** as the BER that achieves stable inflation following the redistribution of workers from the inflating private sector to the fixed price JG sector.

The NAIBER is a full employment steady-state JG level which is dependent on a range of factors, including the historical path the economy has taken.

An aim of government is to minimise the NAIBER so that higher levels of non-JG employment can be sustained with stable inflation. Initiatives that may reduce the value of the NAIBER include public education to stimulate skill development and engender high productivity growth, institutionalised wage setting processes where productivity growth is shared equitably across all income claimants, and restrictions on anti-competitive cartels which should reduce pressures for profit margin push.

However, while central banks and treasuries devote a lot of resources in trying to estimate the NAIRU, we consider it would not be worth trying to estimate or target a particular NAIBER. The point is that the aim of policy is to fully employ labour while maintaining price stability.

Open economy impacts

The JG requires a flexible exchange rate if it is to be effective. A one-off increase in import spending is likely to occur when the policy is introduced because the JG workers will have higher disposable incomes than before.

In most nations, the impact would be modest. We would expect any depreciation in the exchange rate to have low exchange rate pass-through effects on the price level via higher import prices and to provide a modest boost to net exports and local employment, as will be explained in Chapter 24.

Would the NAIBER be higher than the NAIRU?

We have learned that the NAIRU defines the unemployment buffer stock associated with stable inflation whereas in the employment buffer stock approach to price stability, the **NAIBER** is the BER that results in stable inflation via the redistribution of workers from the inflating private sector to the fixed price JG sector.

The main principle of a buffer stock scheme like the JG is straightforward. It buys off the bottom at zero bid, which means that the worker has no other employer bidding for their services and cannot put pressure on wages that are above this floor. The choice of the wage floor (the JG wage) may have a one-off effect on the price level, but does not cause wages or prices to continue to rise.

An interesting question to explore relates to the relative sizes of the NAIBER *vis à vis* the NAIRU. There are two arguments that might be used to argue that the NAIBER would have to be larger than the NAIRU for an equivalent amount of inflation control.

First, the intuitive but somewhat inexact view is that because JG workers will have higher incomes (than when they were unemployed), a switch to this policy would always see demand levels higher than under a NAIRU world.

As a matter of logic then, if the NAIRU achieved output levels commensurate with price stability, other things being equal, a higher demand level maintained through NAIBER would have to generate inflationary impulses. So according to this view, the level of unemployment associated with the NAIRU is intrinsically tied to a unique level of demand at which inflation stabilises.

It should be noted that while it is clear that JG workers will enjoy higher purchasing power under a JG compared to their outcomes under a NAIRU policy, it is not inevitable that aggregate demand overall would rise with the introduction of JG. Indeed, government could engage in a limited austerity policy to lower aggregate demand (cutting non-JG spending and raising tax rates) which would lower aggregate demand but could not increase unemployment as workers shed by the non-government sector would find jobs in the JG. While it is not necessary and probably not even desired that government do this, it is a policy option that is available should government fear that a JG programme would otherwise increase aggregate demand excessively.

But assuming aggregate demand is higher when the JG is introduced than that which prevailed in the NAIRU economy, we might wonder why higher inflation is not inevitable as we replace unemployment with (higher paying) employment. Some worry that the budget deficit will be larger with the JG, but that does not necessarily invoke inflationary pressures because by definition, if that occurs, the deficit is satisfying a net savings desire by the private domestic sector (accumulated as net financial claims on government). That is, the extra government spending on a JG programme results in a deficit that is matched by the non-government sector's net saving (and accumulation of net financial claims on government).

Additionally, in demand-constrained economies, firms are likely to increase capacity utilisation to meet the higher sales volumes rather than risk losing market share by increasing prices. There would be no obvious cost pressures forcing firms to increase prices.

Further, the aggregate demand impulse required to return the economy to what we might call loose full employment under the JG is less than would be required in a NAIRU economy, where the government would have to pay market prices to bring the idle resources back into productive use. In that context, it is clear that if there was any demand pull inflation, it would be lower under the JG. So, there are no new problems faced by employers who wish to hire labour to meet the higher sales levels.

Finally, any initial rise in demand will stimulate private sector employment growth while reducing JG employment and spending.

The second argument, which is related to the first point, claims that the introduction of the JG reduces the threat of unemployment which serves to discipline the wage setting process. In the NAIRU logic, workers may consider the JG to be a better option than unemployment because it reduces the fear of job loss. Without the threat of unemployment, wage-bargaining workers then may have less incentive to moderate their wage demands notwithstanding the likely disciplining role of wait unemployment in skilled labour markets. However, the impact on the price level by the introduction of the JG will depend in part, on qualitative aspects of the JG pool relative to the NAIRU unemployment buffer.

The functioning and effectiveness of the buffer stock in question is critical to its operation as a price anchor. In an economy that uses unemployment buffer stocks to discipline the inflation process, there is overwhelming evidence that long-term unemployment generates costs far in excess of the lost output that is sacrificed every day that the economy is not at full employment.

It is clear that the more immediately employable are the unemployed, the better the unemployment price anchor will function. After an extended downturn, the unemployment buffer stock will be composed of a significant proportion of long-term unemployed, who are not effective as a reserve army of the unemployed.

JG workers are far more likely to have retained higher levels of relevant skills than those who are forced to succumb to lengthy spells of unemployment. It is thus reasonable to assume that an employer would consider a JG worker, who is already demonstrating a commitment to working, to be a superior training prospect relative to an unemployed and/or hidden unemployed worker.

The JG policy would thus reduce the hysteretic inertia embodied in the long-term unemployed and allow for a smoother private sector expansion. Therefore, JG workers would constitute a more credible threat to the current private sector employees than say, the long-term unemployed.

When wage pressures mount, an employer would be more likely to exercise resistance if they knew they could hire from the fixed price JG pool. This changes the bargaining environment rather significantly because firms now have reduced hiring costs. Previously, the same firms would have lowered their hiring standards and provided on-the-job training and vestibule training (for example, controlled skill development outside the production area) as the labour market tightened.

As a consequence, longer-term planning with cost control would be enhanced. So, in this sense, the inflation restraint exerted via the NAIBER is likely to be more effective than using a NAIRU strategy.

In summary, the JG buffer stock is likely to be a qualitatively superior inflation-fighting pool than the unemployed stock under a NAIRU. In that sense, the NAIBER will be lower than the NAIRU, which means that private sector employment can be higher before the inflation barrier is reached.

Another associated factor relates to the behaviour of professional occupational markets. In those markets, while any wait unemployment will discipline wage demands, the demand pressures may eventually exhaust this stock and wage-price pressures may develop.

With a strong and responsive tertiary education sector, combined with strong firm training processes, skill bottlenecks can be avoided more readily under the JG than with an unemployed buffer stock in place. The JG workers would already be maintaining their general skills as a consequence of an ongoing attachment to the employed workforce. The qualitative aspects of the unemployed pool deteriorate with duration of unemployment, making their transition back into the workforce more problematic. Thus, the long-term unemployed exert very little downward pressure on wages growth because they are a less credible substitute.

Employment buffer stocks and responsible fiscal design

In an open economy, the level of economic activity (output) that is determined by private domestic spending (consumption plus investment) and net external spending (exports minus imports) might not be sufficient to generate full employment. Further, if one or more of those components of spending declines, then activity will decline.

We have already defined a **spending gap** as the spending required to create sufficient demand to elicit an output level which, at current levels of productivity, will provide enough jobs (measured in working hours) for all the workers who desire to work.

A zero spending gap occurs when there is full employment. We assume that there is no capacity-constrained unemployment where the level of capital stock is unable to support enough jobs to satisfy the available labour supply at existing productivity levels.

In Chapter 22, we introduce the **full employment fiscal deficit condition**, which recognises that a currency-issuing government should ensure that there is no spending gap which would cause the economy to depart from full employment.

We show that because monetary policy changes are relatively ineffective as a counter-stabilisation policy tool, then if private spending declines from a given position of full employment, the only way that the spending gap can be filled is via a fiscal stimulus, directly through government spending and/or indirectly via a tax cut, which will increase private disposable income and stimulate subsequent private spending.

The basic fiscal rule that government must satisfy to achieve full employment is that the **discretionary fiscal position (deficit or surplus) must fill the gap between the savings minus investment minus the gap between exports net of imports at their full employment values** (see Chapter 22 for more detail). If the fiscal deficit is not sufficient, then national income will fall and full employment will not be achieved. If the government tries to expand the fiscal deficit beyond the full employment limit, then nominal spending will outstrip the capacity of the economy to respond by increasing output, and while nominal income will rise, it will be due to price effects (that is, inflation would occur).

In this sense, MMT specifies a strict discipline on fiscal policy. If the goal is full employment and price stability, then the full employment fiscal position condition must be met.

The question then arises: how do employment buffer stocks relate to this condition?

We used the term **loose full employment** in relation to the JG because the employment generated is at minimum wages. The government expands the JG pool by **purchasing off the bottom** of the labour market. In that context, the automatic stabiliser response associated with the conduct of the JG represents the minimum fiscal shift that is required to maintain employment at its previous level in the face of a falling level of private demand.

The maintenance of the level of employment, however, is achieved by raising the BER. That is, more workers are working at the programme's set wage and less above that at market wages when the JG pool expands.

The government may decide that it has non-inflationary room to then expand non-JG employment via direct job creation in the career section of the public sector or by a general fiscal stimulus designed to increase private sector employment. In this case, the actual deficit spending that will satisfy the full employment fiscal deficit condition varies according to the proportion of the deficit that is associated with JG employment.

A plausible adjustment path

A plausible story to show the dynamics of a JG economy compared to a NAIRU economy would begin with an economy with two labour sub-markets: Sector A (primary) and Sector B (secondary) which broadly correspond to the dual labour market depictions in the labour economics literature. A distinction is made between stable, well-paid primary jobs and low-paid, precarious secondary jobs. Assume as before that firms set prices according to mark-ups on unit costs in each sector.

Wage setting in Sector A is contractual and responds in an inverse and lagged fashion to the ratio of sector wages (Sector A/Sector B) and to the wait unemployment level that consists of displaced Sector A workers who think they will be re-employed soon in Sector A.

Thus, when the ratio of Sector A wages to Sector B wages falls, workers in Sector A will eventually seek to reinstate the past relativity, which reflects their sense of worth in the wage structure and their bargaining capacity as skilled workers. Increasing numbers of workers who are waiting for work in Sector A instead of taking Sector B jobs (wait unemployment), also depress wages growth in Sector A.

Consider a carefully designed government stimulus that immediately increases output and employment in both these sectors of a NAIRU economy. Wages are relatively flexible upwards in Sector B and respond immediately. The compression of the Sector A/Sector B wage relativity stimulates wage growth in Sector A after a time. Wait unemployment falls due to the rising employment demand in Sector A, but also rises due to the increased probability of getting a job in Sector A. That is, workers who had previously taken Sector B jobs in desperation, or were classified as being outside the labour force, may leave their Sector B jobs or re-enter the labour force in the expectation of obtaining a better-paying Sector A job which is more in line with their skill levels. The net effect of these two movements is unclear at the conceptual level. The total unemployment rate falls after participation effects are absorbed. The wage growth in both sectors may force firms to increase prices, although this will be attenuated somewhat by rising productivity as utilisation increases.

A combination of wage-wage and wage-price mechanisms in a soft product market can then drive inflation. These are the type of adjustments that are described in a Phillips curve economy.

To stop inflation, the government must supress demand. The higher unemployment brings the real income expectations of workers and firms into line with the available real income and the inflation stabilises. This is a typical NAIRU story.

Now consider what would be different in a JG economy. Introducing the JG policy into the depressed economy puts pressure on Sector B employers to restructure their jobs in order to maintain their workforces.

For given productivity levels, the JG wage constitutes a floor in the economy's cost structure. The dynamics of this economy change significantly.

The elimination of all but wait unemployment in Sector A and frictional unemployment does not distort the relative wage structure, so that the wage-wage pressures arising from variations in the Sector A/Sector B relativity, which were prominent previously, are now reduced.

The wages of JG workers (and hence their spending) represents a modest increment to nominal demand, given that the state was typically already supporting them on unemployment and other social benefits. It is possible that the rising aggregate demand tightens the product market, and the demand for labour rises in Sector A.

But there are no new problems faced by employers who wish to hire labour to meet the higher sales levels in this environment. They must pay the going wage rate, which is still preferred to the lower JG wage by the appropriately skilled workers. The rising aggregate demand *per se* does not invoke inflationary pressures if firms can increase capacity utilisation to meet the higher sales volumes.

With respect to the behaviour of workers in Sector B, one might think that the provision of the JG would lead to workers quitting bad private sector employers. It is clear that with a JG, wage bargaining in Sector B is freed from the general threat of unemployment. However, it is unclear whether this will lead to higher wage demands than otherwise. In professional occupational markets, some wait unemployment will remain. Skilled workers who are laid off are likely to receive payouts that forestall their need to get immediate work. They have a disincentive to immediately take a JG job, which is a low wage and possibly stigmatised option. Wait unemployment disciplines wage demands in Sector A. However, demand pressures may eventually exhaust this stock, and wage-price pressures may develop.

As noted earlier, a crucial point is that the JG does not rely on the government spending at market prices which then exploits the expenditure multiplier to achieve full employment, as is characteristic of traditional Keynesian pump priming. In this sense, traditional Keynesian remedies fail to provide an integrated full employment price anchor.

19.6 Impact on the Phillips Curve

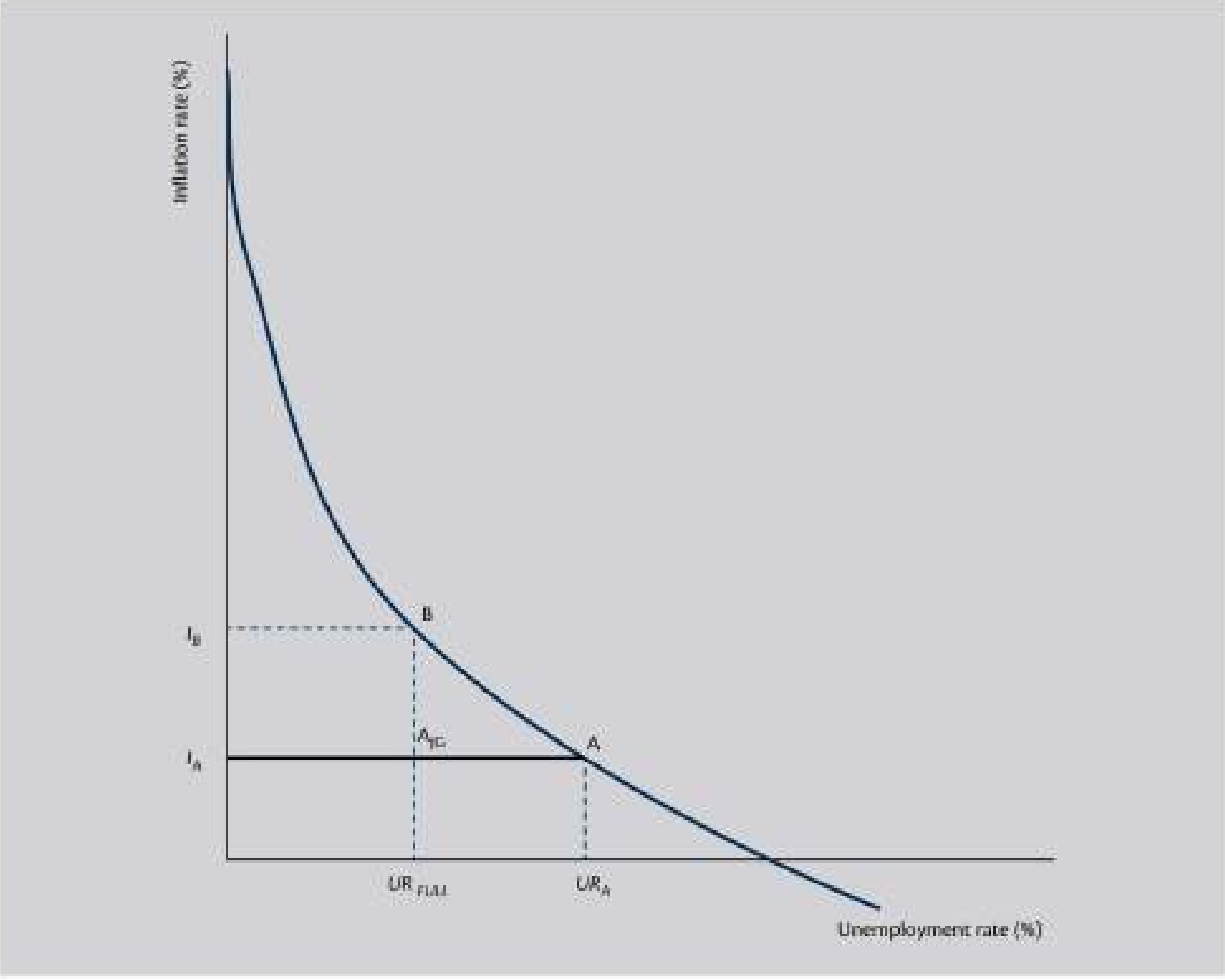
We will now examine the impact of the introduction of a JG on the traditional Phillips curve trade-off.

Consider Figure 19.3. In a Phillips curve world, imagine that the unemployment rate was at UR_A and the inflation rate was I_A. The full employment unemployment rate is UR_{FULL}, which denotes frictional unemployment.

The government is under pressure to reduce the excessive unemployment. If it increased aggregate demand, wage-wage and wage-price pressures would drive the inflation rate up to I_B (a movement along the Phillips curve from Point A to Point B) and achieve full employment.

However, there is no guarantee that the inflation rate would remain stable at I_B. Certainly, the NAIRU model would predict that bargaining agents would incorporate the new higher inflation rate into their expectations and

Figure 19.3 The JG and the Phillips curve



the Phillips curve would start moving out. Whether that happens in reality is not relevant here (we considered those issues in Chapter 18).

If the government initially responded to the excessive unemployment at Point A by introducing a JG it could absorb workers in jobs commensurate with the difference between UR_A and UR_{FULL}, although in reality as more work became available, workers from outside the labour force (the hidden unemployed) would also take JG jobs in preference to remaining without income.

But whatever the quantum of workers that would initially be absorbed in the JG pool, the economy would move from A to A_{JG} rather than from A to B.

> **In other words, the introduction of the JG eliminates the Phillips curve. The macroeconomic opportunities facing the government are not dictated by a perceived unemployment and inflation trade-off which might be unstable (as in a NAIRU world).**

Rather, full employment and price stability go hand in hand.

Benjamin Graham wrote in the 1930s about the idea of stabilising prices and standards of living by surplus storage. He documents how a government might deal with surplus production in the economy: "[The] State may deal with actual or threatened surplus in one of four ways: (a) by preventing it; (b) by destroying it; (c) by dumping it; or (d) by conserving it." Graham (1937: 18). In the context of an excess supply of labour, governments now tend to choose the dumping strategy via the unemployment buffer stock approach (the NAIRU). However, it is less wasteful to use the conservation approach, which is reflected in the JG framework. The JG approach is also based on the maintenance of a variable buffer stock of jobs in line with fluctuations in private demand. However, the weaknesses of the agricultural scheme do not apply to a JG. First, if there is a price guarantee (the JG wage) below the prevailing market price, and a buffer stock of working hours which is designed to absorb the excess supply of labour at the JG wage, then a form of full employment can be generated without tinkering with the price structure.

Second, the incentives to overproduce in commodity buffer stock systems do not apply to maintenance of a labour buffer stock because no one is concerned that employed workers would have more children than unemployed workers.

BOX 19.1 BUFFER STOCKS IN AGRICULTURE

The JG bears many similarities to (and a significant difference from) agricultural price support buffer stock schemes that governments have regularly used to stabilise prices and incomes in the agricultural sector.

For example, in November 1970, the Australian government introduced the Wool Floor Price Scheme after hearing submissions from the Wool Council of Australia and the Australian Wool Corporation (AWC). The government set a floor price for wool.

The aim of the system was to stabilise farm incomes and it led to an agreed price for wool being paid to the farmers. The government then stabilised the price at this guaranteed level by using the AWC to purchase stocks of wool in the auction markets if demand was low and selling it if demand was high.

By being prepared to hold buffer stocks of wool in times of low demand and to resell them in times of high demand, the government was able to guarantee incomes for the farmers around the stable price.

The programme thus stabilised wool prices and farmer incomes, and helped to stabilise the consumption of these farmers. In all these ways, the buffer stock programme was stabilising.

Note that the floor price was not inflationary because it simply guaranteed that wool prices could not fall below the floor.

The contention that ultimately led to the demise of the system was whether the guarantee constituted a reasonable level of output in a time of declining demand. Farmers clearly had an incentive to overproduce wool knowing that the government would buy any excess not demanded by the auction markets at the floor price.

Conclusion

This chapter has contrasted the macroeconomic outcomes associated with the implementation of unemployment and employment buffer stock policies which are designed to achieve full employment and price stability. It has demonstrated that only an employment buffer stock policy can achieve these macroeconomic targets.

There are many microeconomic factors that are relevant to a full understanding of how a JG would work in practice. There are questions relating to the type of jobs, the levels of government involved in funding and operations, the relationship with the existing income support system, the integration of training pathways into the policy, the role of trade unions, the choices available to workers for fractional employment, the capacity of the government to sack workers and more.

While these are important factors, which have been dealt with in the literature, they lie outside of our macroeconomic focus in this textbook. More information and analysis can be found in the references that follow.

References

Beveridge, W. (1945) "Full Employment in a Free Society: A Summary", *New Statesman and Nation and Reynolds News*, London. Available at: http://lib-161.lse.ac.uk/archives/beveridge/9A_79_Full_employment_in_a_free_society.pdf, accessed 25 February 2016.

Forstater, M. (2006) "Working for a Better World: The Social and Economic Benefits of Employment Guarantee Schemes", Center for Full Employment and Price Stability, University of Missouri-Kansas City and the Levy Economics Institute of Bard College, available at: http://www.economistsforfullemployment.org/knowledge/presentations/Session1_Forstater.pdf, accessed 2 May 2017.

Graham, B. (1937) *Storage and Stability*, New York: McGraw Hill.

Kaboub (2008) "Employer of Last Resort Schemes" in P.A. O'Hara (ed.), *International Encyclopedia of Public Policy: Governance in a Global Age*, Perth: GPERU.

Minsky, H. (1986) *Stabilizing an Unstable Economy*, New Haven, CT: Yale University Press.

Mitchell, W.F. and Muysken, J. (2008) *Full Employment Abandoned*, Cheltenham and Northampton, MA: Edward Elgar.

Modigliani, F. (2000) "Europe's Economic Problems", *Carpe Oeconomiam Papers in Economics*, 3rd Monetary and Finance Lecture, Freiburg, 6 April.

Ormerod, P. (1994) *The Death of Economics*, London: Faber & Faber.

Petty, W. (1662) "A Treatise of Taxes and Contributions", in Murphy, A.E. (ed.) (1997) Monetary Theory 1601–1758, Vol. 2, London and New York: Routledge.

Pierson, John H.G. (1980) *Full Employment without Inflation: Essays on the Economic Performance Insurance (EPI) Proposal*. Montclair, NJ: Allanheld, Osmun & Co.

Sen, A. (1997) "Inequality, Unemployment and Contemporary Europe." *International Labour Review*, 136(2), 155–172.

Wernette, J.P. (1945) *Financing Full Employment*. Cambridge, MA: Harvard University Press.

Endnotes

1. Strictly speaking the USA's central bank, the Fed, does not target inflation, which would not be consistent with its dual mandate. However, as it has come to see low inflation as a precondition to robust GDP and employment growth, it puts a lot of weight on pursuit of relative price stability.
2. In view of the significance of underemployment in moderating wage inflation, it would be necessary to respecify the so-called NAIRU.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

PART E

ECONOMIC POLICY IN AN OPEN ECONOMY

CHAPTER

20 INTRODUCTION TO MONETARY AND FISCAL POLICY OPERATIONS

Chapter Outline

Learning Objectives

- Understand the roles of the treasury and the central bank.
- Recognise why and how liquidity management by the central bank accompanies the operation of fiscal policy.
- Acknowledge that the design of the taxation system should be motivated by equity and behavioural objectives and not revenue raising.
- Appreciate that a necessary condition for the independence of macroeconomic policy is the sovereignty of the domestic currency.

20.1 Introduction

In this chapter, we will address three main topics:

1. How is fiscal policy conducted?
2. How is monetary policy conducted?
3. How do the central bank and the treasury coordinate their operations to enable the sovereign government to spend?

We start this chapter by briefly exploring the roles of the central bank and treasury in an economy with its own floating sovereign currency. We again highlight that while accounting principles are universal, their application to households and firms as currency users differs fundamentally from the central bank as a currency issuer.

Our focus in this chapter is operational practice. For many years there has been a disconnect between the textbooks' treatment of the implementation of fiscal policy and the institutional arrangements in different countries, such as the USA, UK and Australia, which while following similar principles, do differ in the way they conduct fiscal policy. We provide a generic and simplified description of the conduct of fiscal policy.

We also explain why the self-imposed constraints that the political system puts on central banks buying treasury debt directly in the primary market do not affect the operational outcomes in any meaningful way (see Lavoie, 2013; and Tymoigne and Wray, 2013).

We then revisit the role of taxation in a modern monetary system. The chapter concludes with a further discussion of the crucial importance of a sovereign floating currency for policy independence. In the Appendix, we extend the analysis of central bank operations to take account of an open economy.

20.2 The Central Bank

Modern governments operate with a central bank. In some nations, the central bank is formally independent of the treasury although typically the elected government still appoints the senior management ('the board') and maintains the right to overrule monetary policy decisions. But that political control aside, most central bank managers (the Board of Governors, in the case of the US; and the Monetary Policy Committee in the UK) have some independence from elected representatives and from the bureaucracy.

The central bank is empowered to set a target overnight or interbank interest rate, which is now the primary tool for the implementation of monetary policy. The presumption is that an independent and – allegedly – apolitical body charged with the formulation of monetary policy will make better decisions.

In practice, the central bank's independence is not great for a variety of reasons. For example, in the US, the Federal Reserve Bank (known as the 'Fed') is a 'creature of Congress', subject to the laws that Congress has promulgated. Indeed, the Fed was created by an Act of Congress (the 1913 Federal Reserve Act), and Congress has periodically mandated changes to Fed operations. Similarly in the UK and Australia, the Bank of England and Reserve Bank of Australia (RBA) are subject to legislation passed in their respective parliaments. In Australia, the elected government appoints the Governor and the management board of the RBA, and the Treasurer can also reject the RBA board's decisions to change interest rates.

There is a more important reason to doubt the central bank's independence: its interest rate (liquidity) management operations are largely accommodative because it not only responds to the needs of private banks but also to the actions taken by the treasury, which necessitates close coordination with treasury. In this section we will provide a summary of the central bank's operations. In Chapter 23, we will go into more detail by addressing issues related to control of private banking operations, setting of interest rates, and provision of reserves through lender of last resort operations.

Most central banks pursue what has been termed **inflation targeting** because many economists claim that it creates an environment of low and stable inflation, which provides more certainty for the private sector to make spending decisions. Accordingly, the Bank of England is subject to a CPI inflation target of two per cent per year, whereas in Australia, the CPI inflation target has been set between two and three per cent per year since 1993. The US Fed does not set a specific inflation rate target, but the Federal Open Market Committee (FOMC, 2016) states that a CPI inflation rate target of two per cent is "most consistent over the longer run with the Federal Reserve's statutory mandate."[1]

However, there is no hard evidence available to support the claims that inflation targeting improves the functioning of the economy. It is certainly hard to distinguish the outcomes of a targeting nation from non-inflation targeting countries, especially those that have adopted the broader 'fight inflation first' monetary stance, such as the US (see Ball and Sheridan, 2003).

The main problem is that the move to inflation targeting (formal or informal) has prioritised monetary policy as the main macroeconomic policy tool. As a result, governments now see fiscal policy as a passive policy tool and tend to run unduly restrictive fiscal policy stances so as not to contradict the monetary policy stance. This has led to slower average real GDP growth rates and higher unemployment in many nations.

The payments system, reserves and the interbank market

Most central banks pursue their monetary policy stance by announcing a target overnight (interbank) interest rate. In practice, central banks often have a strategy of maintaining the overnight rate within a narrow range around the target. Central banks use several different interest rates to achieve this:

The **interbank rate** (called the **Fed funds rate** in the USA) is the rate at which banks lend reserves overnight to each other to alleviate shortages that might arise.

The **discount rate** is the rate the central bank charges when it loans reserves to banks.

The **deposit rate** is the interest rate paid on reserves held on deposit at the central bank.

In general, the central bank's main target is the interbank rate. It does not set this rate directly but rather uses the discount rate and interest rate it pays on reserves to push or pull the interbank rate to the target. As we explain in more detail in Chapter 23, the deposit rate sets a floor to the interbank rate. This is because banks can always earn that rate on reserves they hold at the central bank, so they will not lend reserves to other banks at a lower rate. The discount rate sets a ceiling to the interbank rate because banks can borrow reserves at the central bank and so will not borrow from other banks at a higher rate. Thus, the interbank rate will tend to fluctuate within the range set by the floor (deposit rate) and ceiling (discount rate). By narrowing the range, the central bank reduces the fluctuation of the interbank rate.

As noted in Chapter 10, private banks hold reserves at the central bank to enable the payments system to function efficiently. Payments for goods and services by customers to retailers using different banks are resolved not only by the respective bank deposit accounts of these buyers and sellers being adjusted, but also by reserves being transferred from the banks of the customers to the banks of the retailers. The operation of the payments system can leave some banks with a shortage of reserves, whereas other banks will have excess reserves. The deficit banks will seek to borrow reserves in the interbank market from the surplus banks and will pay the prevailing interbank interest rate.

It is important to understand that banks do not loan out reserves to customers, contrary to the impression given by orthodox theories of money. They are used exclusively to ensure an efficient payment settlements system between banks and the central bank.

If there are excess reserves overall, then market forces will drive the interbank interest rate toward zero because banks will reduce the price at which they are willing to loan reserves. In this case, the central bank can relieve the downward pressure by removing excess reserves. Likewise, in the case of a shortage of overall reserves, the interbank interest rate will be driven toward the ceiling of its target level. The central bank relieves the upward pressure by supplying more reserves.

Reserves are added through **discount window loans**[2] offered by the central bank (at the 'discount rate') through open market purchases by the central bank of government bonds, and through purchases of gold, foreign currencies, or even private sector financial assets. In other words, banks can secure additional reserves when there is a (banking) system shortage, either through borrowing at the central bank's discount window, or by selling assets to the central bank. In either case, the central bank credits the reserve deposit accounts of the banks at the central bank, which addresses the system shortage.

The central bank reverses these actions in the case of excess reserves in the banking system, that is, when the banks have more reserves than they wish to hold (or are required to hold). Banks with excess reserves can pay down loans at the discount window, or they can buy assets (usually treasury debt, although possibly foreign currency or private assets) from the central bank. The central bank will then debit their reserve accounts.

The central bank must predict reserve supplies and demands on a daily basis. Fortunately, it is easy to determine whether the banking system faces excess or deficient reserves. The overnight rate will move away from its target, triggering a nearly automatic offsetting addition or drain of reserves by the central bank.[3]

In normal times, central banks accommodate the private bank demand for reserves so that they can control the overnight interest rate. This quantity of reserves is non-discretionary from the point of view of the central bank. It is the interest rate target that is discretionary. In a crisis, the demand for central bank reserves can rise suddenly because banks will be reluctant to provide loans to each other. In these situations, the central bank must step in to supply the additional reserves.

The central bank needs to have good lines of communication with treasury to ensure that the financial system is not disrupted by the treasury running either surpluses or deficits, both of which impact on bank reserves. In Sections 20.3 and 20.4, we outline the operation of fiscal policy and how it impacts on reserves.

Finally, we need to recognise that central banks perform other functions, including acting as a lender of last resort. For example, a bank in financial difficulty may not be able to borrow reserves in the interbank lending market, even if excess reserves exist at the aggregate level. This is because private banks do not want to lend to troubled banks, fearing there could be repayment problems. The central bank will lend to the troubled bank and will step in to resolve the problems or to shut down the bank if needed.

Central banks also regulate and supervise private banks and other financial institutions. For example, the central bank might prohibit banks from making certain kinds of loans (that is, credit controls) or from issuing some kinds of deposits. In many nations, the central bank plays some role in ensuring the 'safety and soundness' of individual banks as well as of the financial system as a whole. Such roles are also performed by other bodies: divisions of the treasury, state or provincial government offices, and even independent financial sector regulators. In addition, many nations enforce international guidelines on the behaviour of financial institutions, such as those adopted in the Basel Accords, which are designed to ensure financial stability.

A detailed examination of bank regulation and supervision is beyond the scope of a macroeconomics textbook. We will however discuss these matters briefly later in the text, when we discuss financial instability and global financial crises.

20.3 The Treasury

The treasury is the fiscal agent of the elected government in that it operationalises fiscal policy via the instruments of government spending and taxation. Note that in some nations the treasury function is performed by the Ministry of Finance. We will use the generic term treasury, which is common in Anglo nations.

In the distant past, the treasury would spend directly through the issue of money-denominated IOUs, whether these were tally sticks, metallic coins, or paper money. The administrative branch would spend up to the specified amount through treasury issue of new money, with provisions being made for overrides.

The treasury would also be responsible for collecting taxes in the form approved by elected representatives. Normally, this would include the money-denominated IOUs issued by treasury in its spending. In addition, however, the treasury would sometimes be permitted to accept other IOUs, including currency issued by other nations, other types of domestic government IOUs, or even some types of privately issued IOUs denominated in the domestic currency. Modern treasuries only accept the liabilities of their own governments, mostly in the form of central bank reserves and notes, and sometimes treasury coins or notes.

Government and private financial accounting

Even though some principles of accounting are universal, federal financial accounting has never followed, and should not follow, the procedures adopted by households or business firms. We outlined the arguments behind this assertion in Chapter 2 and summarise and supplement these arguments here.

First, the government's objective should be the pursuit of public purpose, that is, general welfare. There is no necessary correlation between this objective and the achievement of a fiscal surplus or deficit, or higher or lower indebtedness.

Second, the government is sovereign. This fact gives to the government an authority that households and firms do not have. Government has the power to tax and to issue money. The power to tax means that government does not need to sell products or otherwise 'earn' income like firms or households. The power to issue currency means that it can make purchases by dispensing its own IOUs. In short, **currency-issuing governments like**

those in Britain, the United States, Japan and Australia can never run out of money. These governments can purchase whatever goods and services are for sale in the currency they issue. They must consider how best to deploy the real resources available to the economy, but there are no intrinsic 'financial' constraints that are imposed on a currency-issuing government.

While it is common to regard government tax revenue as income, this revenue is not comparable to that of firms or households. Government can choose to impose new taxes or raise tax rates.

There is no operational procedure through which the national government uses tax receipts or borrowings for its spending. If households choose to pay taxes in cash, the treasury simply issues a receipt and can choose to shred the cash. It does not need these tax receipts in order to spend. Thus, it is a mistake to look at the national government's tax receipts as an equivalent concept to the income of households or firms.

Also, fiscal surpluses (taxation revenue greater than government spending) today do not provide governments with a greater capacity to meet future spending needs, nor do fiscal deficits (taxation revenue less than government spending) erode that capacity.

Indeed, there is no evidence, nor any economic theory, behind the proposition that national government spending ever needs to match national government tax receipts over any period, short or long. The deficit per unit of time is simply the difference between taxing and spending over that time. It is not an indication that the government is 'running out of money' or 'spending beyond its means'. The size of the deficit does not provide evidence that the government is 'spending too much' or 'taxing too little'. A large deficit could be consistent with both 'spending too little' and 'taxing too much'.

Taxing, on the one hand, and spending, on the other, are operationally independent procedures.

On the other hand, private firms are revenue-constrained because they cannot force buyers to purchase their products or their debt. Even firms with market power recognise that consumers will find substitutes if prices are raised too much, and that lenders will cut off loans to firms with too much debt. Likewise, households cannot force anyone to give them more income, or to lend more to them. Their spending is constrained by income and previous saving plus the ability to borrow.

The government is in an entirely different situation. Taxation creates a demand for public spending to make available the currency required to pay the taxes. No private firm can generate demand for its output in this way. Neither firms nor households can live beyond their financial means indefinitely by accumulating debt. Eventually they must sacrifice spending to pay the debts back. Thus, firms, households, and even state and local governments, require income or accumulated savings or need to borrow in order to spend.

These statements are not controversial. They are matters of fact. Nor should they be construed as implying that government should raise taxes sky high or spend without limit. However, they do imply that financing national government spending is different to private budgeting which is necessary to plan future expenditures.

MMT teaches us that experience in managing a household budget provides no guidance about the management of the national government fiscal position. Yet, on a daily basis, the message delivered by the media and most politicians is that the same principles apply.

Sectoral balances

The difference between microeconomic and macroeconomic accounting is also pertinent. An individual household or firm has a balance sheet that consists of assets and liabilities. The spending of that household or firm is constrained by its income and balance sheet, that is, by its ability to sell assets or to borrow against them. A household or firm must get the approval of a bank before its spending can exceed income (unless it has financial[4] or physical assets to sell), and therefore borrowing is subject to banking norms.

On the other hand, if we consider households and firms in aggregate, the situation is different. The private domestic sector's ability to deficit spend (that is, to spend more than its income), depends on the willingness of another sector (the government and/or foreign sector) to spend less than its income. **For one sector to run a deficit, another must run a surplus** (as this textbook has emphasised, see Chapter 6). This surplus is saving, which constitutes claims against the deficit sector. In principle, there is no reason why one sector cannot run perpetual deficits, so long as at least one other sector wants to run surpluses.

In the real world, we observe that most national governments (including those in the US, the UK and Australia) tend to run persistent deficits. This is matched by a tendency of the non-government sector, which comprises the private domestic and external sectors, to save overall, that is, spend less than its total income. The non-government sector accumulates net claims on the government; **the non-government sector's overall saving is equal (by identity) to the government's deficits** over any given period.

At the same time, the non-government sector's net accumulation of financial assets (or **net financial wealth**) exactly equals the government's total net issue of debt from the inception of the nation. Debt issued between private parties cancels out in net terms for the private sector as a whole; but that between the government and the non-government sector remains, with the non-government sector's net financial wealth consisting of the government's net debt.

This identity does not change once we allow for a foreign sector, which is just a part of the non-government sector (we divide the non-government sector into its two components: the domestic private sector and the foreign sector). Since nations such as the US, the UK and Australia have in recent decades run persistent current account deficits, the foreign sector[5] has been accumulating net financial claims on those nations in the currencies issued by each. These are initially held in the form of cash or reserve balances at the respective central banks of countries with external surpluses, but they are then typically exchanged for government debt of the deficit countries in order to earn interest.

Sectoral balances are linked by an identity, so that by definition government deficits equal non-government surpluses, and government debt equals non-government net financial wealth. Yet, as noted in Chapter 6, these macroeconomic relations are not obvious when one analyses individual firms or households.

20.4 Coordination of Monetary and Fiscal Operations

A consolidated government, consisting of a treasury and a central bank, was typically the model that was adopted at the formation of most modern states. Even today, there are many countries that operate without a clear division of responsibilities between the central bank and treasury. Hence, the consolidated government is of some theoretical interest.

While it has been common in the MMT literature to begin with an analysis that **consolidates the central bank and treasury into a sovereign government**, we will maintain the division of responsibilities between the central bank and treasury.

Duties of the central bank

- Issues currency notes in many nations and in rare cases also issues coins.
- Issues reserves (at discount window and through open market purchases of government bonds in secondary markets).
- Sets overnight interest rates, operates a clearing house for interbank payments, and operates clearing for bank settlements with the treasury.
- Conducts other transactions relating to foreign exchange and gold.

The separation of responsibilities between the central bank and treasury leads to one extra function for the central bank, namely acting as an intermediary with respect to payments made to and from private banks to treasury. This arises because private banks hold their reserve deposit accounts at the central bank and do not have accounts at the treasury. Thus, the analytical simplification associated with consolidation of the treasury and the central bank has minimal effect.

Duties of the treasury

- Makes payments to the non-government sector.
- Receives tax payments from the non-government sector.
- Issues new government bonds (usually through a specialised public debt management agency).
- Issues coins (in the US).

Note that today in most countries the treasury makes and receives payments through its account at its central bank, much as you make and receive payments through your bank account.

The two main **voluntary, operational** rules, typical of many countries, are that:

1. The treasury writes cheques on its account at the central bank when engaged in spending. According to the operational rule adopted, it must have sufficient deposits in that account before it writes a cheque.
2. The treasury cannot sell newly issued bonds to the central bank on the primary market; it must sell them to private banks or other investors. However, the central bank can buy these bonds from private banks or other investors on the secondary market.

Thus, the treasury typically is prevented by legislation or other rules from selling bonds to its own bank (the central bank on whose account it draws to spend), but it can sell bonds directly to private banks. It is important to understand that these restrictions are not intrinsic, but are voluntarily imposed by the government on itself.

The reason for these restrictions has more to do with satisfying ideological preferences to make it hard for governments to spend, rather than any economic or financial necessity or sound fiscal practice. We see that in emergencies the restrictions are often quickly relaxed to provide more flexibility to government to use its currency-issuing capacity to meet the challenge of the crisis (for example, during the GFC).

When it spends, the treasury injects currency (normally in the form of central bank reserves) into the economy. Private banks credit the deposit accounts of the sellers of goods and services (or the recipients of transfer payments) at the private banks. At the same time, the central bank credits the private banks' reserves. This means that the banks must have reserve accounts at the central bank to be credited.

The treasury usually accepts only currency in the form of central bank reserves for the payment of taxes. This means that when taxes are paid the bank entries of taxpayers are debited, and the central bank debits the reserves of the taxpayers' banks by an equivalent amount.

Deficit spending means net currency emission; fiscal surpluses mean the stock of net currency (that is, bank reserves and holdings of notes and coins) of the non-government sector is reduced.

The view that a central bank might choose to print money to finance a fiscal deficit is flawed. If the government runs a deficit, it inevitably net credits bank reserve accounts (credits exceed debits) in the first instance. All else being equal, a government deficit will generate excess reserves in the banking system. This places downward pressure on the overnight interest rate, as discussed earlier.

However, if the central bank operates with a positive overnight target interest rate, it must either pay interest on reserves or have an interest paying debt instrument it can use to set a floor to prevent overnight rates from falling below the target due to excess reserves created by deficit spending. Typically, when government spending creates excess reserves, the central bank sells bonds, usually treasury debt, as an attractive interest earning alternative to bank reserves, which can be used to mop up the excess reserves. This is referred to as **open market operations** (OMO).

The purchasing banks are making choices about the composition of their portfolios, taking into account any requirement to hold reserves and their operational need to hold reserves for clearing purposes. If they have extra reserves, they seek to buy interest earning assets and the higher rates offered on treasury debts are attractive. In normal times, banks do not want to hold any excess reserves and open market sales ensure they won't have to do so. To be clear, the central bank's interest in engaging in these operations is to ensure that it hits its overnight interest rate target.

These activities are coordinated with the treasury, which will usually issue new bonds more or less in step with its deficit spending. This is because the central bank would run out of bonds to sell to drain the excess reserves created by deficit spending. The important point however is that such central bank operations are not a choice undertaken to 'finance' budget deficits, but rather are required to ensure that they can consistently achieve their interest rate targets. The quantity of 'liquidity' (reserves) is not normally a discretionary policy because excess reserves will drive overnight rates below the target (potentially all the way to zero) *unless* the central bank pays interest on reserves that is equal to the target rate. So in normal times, the central bank uses open market sales to eliminate excess reserves.[6]

A numerical example using balance sheets

We now provide a simplified analysis in which treasury engages in net government spending of $100.

We assume that the interest rate paid by the central bank on reserves held by the private banks is zero, but the target interbank rate is positive. Treasury has an account at the central bank. It must have sufficient deposits at the central bank to enable its planned spending.

For simplicity and to reflect the notion of the treasury and central bank being consolidated into one entity, we assume that the central bank **can** purchase treasury debt on the primary market.

Table 20.1 shows the balance sheets associated with treasury's net government spending of $100. In Stage 1 treasury increases its deposits at the central bank by selling $100 of treasury debt to the central bank. Thus, we are highlighting the intrinsic features of a modern monetary system in which voluntary constraints are not imposed. In Marxist terminology, we are stripping back the 'veil of ideology'.

When treasury net spends, there is an overall increase in non-government sector bank deposits of $100, representing payments for goods and services sold to the government (Stage 2). At the same time the reserves held by the private banks at the central bank increase by $100, which represents an additional asset for the private banks and a liability for the central bank.

The rise in the liabilities of the private banks, via the rise in deposits of the non-government, is matched by their increased holdings of reserves at the central bank. Thus, the net positions of the central bank and the banks are unchanged. Treasury deposits at the central bank have fallen back to their initial level.

The reserves held by the private banks at the central bank, which enable the operation of the payments system, have risen by $100 and while economic activity has increased, private banks may be reluctant to hold an additional $100 in reserves.

Assume they have a reserve ratio of ten per cent, which means they desire to hold $10 of these extra reserves. Those private banks holding excess reserves will try to lend their excess reserves of $90 to other banks. Given that there is a **system-wide excess**, that is, an overall excess supply of reserves, the interbank rate would be driven down below its target level by this lending activity in the absence of central bank action.

Table 20.1 Balance sheets associated with net government spending ($)

Stage	Central bank				Private banks			
	Assets		Liabilities		Assets		Liabilities	
1	Treasury bonds	+100	Treasury deposits	+100				
2			Treasury deposits	−100			Non-governmental sector deposits	+100
			Bank reserves	+100	Bank reserves	+100		
3	Treasury bonds	−90	Bank reserves	−90	Bank reserves Treasury bonds	−90 +90	Non-governmental sector deposits	
Δ Stocks	Treasury bonds	+10	Bank reserves	+10	Bank reserves	+10	Non-governmental sector deposits	+100
					Treasury bonds	+90		

Note: $ amounts with either a plus or minus sign in front indicate changes to balance sheet items at each step after treasury borrows and spends $100 and the central bank engages in OMO (liquidity management), selling $90 of treasury debt to the private banks. Note that balances are maintained for each entity and across entities at each step.

Source: Adapted from Lavoie (2013:11) with permission from Taylor & Francis Ltd. Column 1 has been added for clarity.

The central bank will offer $90-worth of treasury debt to the private banks, which will attract an interest rate in excess of the target interbank rate (Stage 3). Banks holding the $90 of excess reserves will have the incentive to buy this treasury debt. This action by the central bank will remove the downward pressure on the interbank rate and thus protect the integrity of monetary policy, which is identified with setting a target interbank rate. The overall changes in the balance entries are shown in the final row of Table 20.1 (Δ Stocks).

Thus the coordination of central bank and treasury operations is required to implement this programme of government deficit spending.[7]

In our example, the **monetary base**, which is generally defined as the total bank reserves held by the central bank plus currency held by the non-government sector (that is, banks, non-bank firms and households), has risen by $10 since the non-government sector's holding of notes and coins is unchanged.[8]

The **net financial assets** held by the non-government sector are defined as their holdings of net financial assets plus the monetary base. This has increased by a total of $100, with the private banks' holdings of treasury debt rising by $90 and their holding of cash rising by $10.

Thus, this **vertical transaction** of $100 arising from net spending by treasury in the domestic economy has, as expected, increased the net financial assets of the non-government sector by an equivalent amount (see the distinction between horizontal and vertical transactions in Chapter 6).

We now show the mathematical analysis for the transactions in Table 20.1 in which treasury engages in net government spending of $100. This sum is represented by $(G + iB)$, where G is government spending, i denotes the nominal interest rate, and iB denotes interest payments on existing treasury debt. Net government spending exceeds tax revenue (T), expressed as $(G + iB > T)$. The total change in stock (Δ Stocks) comprises the change in the monetary base (ΔM_h) and the change in the stock of public debt held by the non-government sector (ΔB), so we can write the identity:

$$G + iB - T \equiv \Delta B + \Delta M_h \tag{20.1}$$

A fiscal deficit (left-hand side >0) gives rises to a non-government sector surplus, which takes the form of a change in the stocks of monetary base and treasury debt, held by the non-government sector.

We will see in the next chapter how this identity is also referred to as the 'government budget constraint' and interpreted by mainstream economists as an *ex ante* financial constraint on government, that needs to finance fiscal deficits. In fact, it is just an **accounting identity**, which tells us *ex post* what the changes in the financial aggregates are as a result of discretionary government policy choices and the state of the economy. A currency-issuing government is not financially constrained in the way the mainstream textbooks suggest.

Is there a sufficient demand for treasury debt?

In most developed economies, as a result of the voluntary constraints and accounting conventions that the government places on itself, the treasury must sell bonds if it does not have sufficient deposits at the central bank to cover mandated spending. The important question is whether there will be sufficient demand for them from the banks (and other dealers who are entitled to participate in the primary issuing auctions).

When banks buy bonds from the treasury (via primary auctions), their reserves at the central bank are debited. If a bank that wants to buy bonds has no excess reserves to debit, then it will either go to the interbank market to borrow them from banks with extra reserves, or it will borrow them from the central bank at the discount window (see Chapter 23).

We know that if the banking system has no excess reserves, the central bank will respond to any pressure on overnight interest rates that might be created by banks trying to borrow reserves in order to buy the bonds. It will either lend them at the discount window, or engage in an open market purchase, creating reserves by buying bonds from the non-government sector.

With an interest rate target, the central bank is always accommodating. Thus, banks will always be able to get the reserves they need in order to buy bonds. The banks want the bonds rather than reserves because the interest rate on bonds is higher.

Specialised financial institutions are ready to buy domestic bonds in most countries. For example, in the US there are 21 primary dealers who are obligated to bid at US government debt auctions. Likewise, in the UK treasury bill primary market participants are financial institutions that have agreed, subject to their own due diligence, to bid at UK treasury bill tenders on behalf of investors. These institutions also operate in secondary markets. Essentially, the special dealers are always on standby to purchase new issues of bonds, and then will sell them into secondary markets, or to the central bank that purchases them in open market operations to relieve downward pressure on interest rates when reserves are in short supply.

The evidence reveals that bond issues in the primary auction process are typically oversubscribed. In other words, the demand for new issues of treasury bonds is elastic. Interest rate targeting of the central bank ensures private banks will have the reserves they need to buy the bonds.

20.5 Taxes and Sovereign Spending

Previously we have argued that the imposition of a tax that is payable in the national government's own currency will create demand for that currency. We have also seen that a sovereign government does not need such revenue in its own currency in order to spend. It is somewhat of a misnomer to even refer to these funds as 'revenue', in the sense that private firms, for example, require revenue in order to spend.[9] But the tax flows that national governments receive are certainly not intrinsically necessary to facilitate government spending.

This sounds shocking because we are so accustomed to thinking that taxes pay for government spending. This is true for local governments and states that do not issue the currency. It is also not too far from the truth for nations that adopt a foreign currency or peg their own to gold or foreign currencies.

When a nation pegs, it needs stocks of gold or foreign currency to which it promises to convert its currency on demand. Taxing removes its currency from circulation, making it harder for anyone to present it for redemption in gold or foreign currency. Hence, a prudent practice for a nation that pegs its currency would be to constrain government spending to the level of tax revenue.

But in the case of a government that issues its own sovereign currency without a promise to convert at a fixed value to gold or foreign currency (that is, the government floats its currency), we need to think about the role of taxes in an entirely different way because taxes are not needed to pay for government spending.

Further, the logic is reversed: government must spend (or lend) the currency into the economy before taxpayers can pay taxes in the form of the currency.

Spend first, tax later is the logical sequence.

Some who hear this proposal for the first time may logically jump to the question: 'Well, why not just eliminate taxes altogether?' There are several reasons. First, it is the tax that drives the currency. If we eliminated taxation, people probably would not immediately abandon their use of the currency, but the main driver for its use would be gone.

The second reason to have taxes (once a currency is established and widely adopted) is to reduce aggregate demand. Taxes create **real resource space** in which the government can spend to fulfil its socio-economic mandate. Taxes reduce the non-government sector's purchasing power and hence its ability to command real resources, leaving real resources for the government to command with its spending.

Take a situation where the national government spending is around 30 per cent of GDP, while tax revenue is somewhat less, say 27 per cent. The net injection of spending coming from the national government is thus about three per cent of GDP. If we eliminated taxes (and held all else constant) the net injection rises toward 30 per cent of GDP. That is a huge increase of aggregate demand, and could cause inflation.

Taxes thus free up real resources in the economy (labour and capital) which otherwise would have been used by the non-government sector for private ends. They thus allow the government to spend without coming up against the inflation constraint that would be created once all resources are fully utilised.

Ideally, it is best if tax revenue moves countercyclically, increasing in an expansion and falling in a recession. That helps to make the government's net contribution to the economy countercyclical, which helps to stabilise aggregate demand. In this case the fiscal outcome operates as an automatic stabiliser.

All of this was recognised by Beardsley Ruml, who chaired the US Federal Reserve Bank in the 1940s. He also wrote two important papers on the role of taxes: *Taxes for Revenue are Obsolete* and *Tax Policies for Prosperity* (Ruml, 1946a and 1946b).

Let us first examine his cogent argument that sovereign government does not need taxes for revenue, and then turn to his views on the role of taxes. Ruml emphasised that: "We must recognise that the objective of national fiscal policy is above all to maintain a sound currency and efficient financial institutions; but consistent with the basic purpose, fiscal policy should and can contribute a great deal toward obtaining a high level of productive employment and prosperity" (1946b: 82–3). This view is similar to that being propounded in our textbook.

Ruml also said that the US government gained the ability to pursue these goals after the Second World War due to two changes of great consequence: "The first of these changes is the gaining of vast new experience in the management of central banks. The second change is the elimination, for domestic purposes, of the convertibility of the currency into gold or into any other commodity" (1950: 91). With those two conditions, "[i]t follows that our federal government has final freedom from the money market in meeting its financial requirements ... National states no longer need taxes to get the wherewithal to meet their expenses" (Ruml, 1946b: 84). These insights are applicable to all currency-issuing national governments.

Why then does the national government need taxes? Ruml provides four reasons (1946b: 84):

1. As an instrument of fiscal policy to help stabilise the purchasing power of the dollar;
2. To express public policy in the distribution of wealth and of income as in the case of the progressive income and estate taxes;
3. To express public policy in subsidising or in penalising various industries and economic groups; and
4. To isolate and assess directly the costs of certain national benefits, such as highways and social security.

The first purpose is related to the inflation issue that we discussed above. The second purpose is to use taxes to change the distribution of income and wealth. For example, a progressive tax would reduce income and wealth at the top, while imposing minimal taxes on the poor. The third purpose is to discourage bad behaviour: pollution of air and water, use of tobacco and alcohol, or to make imports more expensive through tariffs (essentially a tax to raise import costs and thereby encourage the purchase of domestic output). The fourth purpose is to allocate the costs of specific public programmes to the beneficiaries. For example, it is common to tax petrol so that those who use the nation's highways will pay for their use (tolls on throughways are another way to do this).

Note that while many would see these taxes as a means to pay for government spending, Ruml (1946a) vehemently denies that view in the title of his article, *Taxes for Revenue are Obsolete*. Government does not need the petrol tax to pay for highways. That tax is designed to make those who will use highways think twice about their support for building them. Government does not need the revenue from a cigarette tax; rather, to improve public health, it wants to raise the cost to those who smoke in order to discourage that behaviour.

The point of these taxes is not the revenue to be generated. Government can always find the money to pay for hospital construction and operation. Rather, it is to reduce the waste of real resources that must be devoted to caring for those who smoke. The ideal cigarette tax would be one that eliminated smoking, not one that maximised revenue to government. Ruml (1946b: 84) said: "The public purpose which is served ... [by the tax] should never be obscured in a tax program under the mask of raising revenue."

Ruml ended both of his 1946 articles by arguing that once we understand what taxes are for, then we can go about ensuring that the overall tax revenue is at the right level. He concluded (1946b: 85):

> *Briefly the idea behind our tax policy should be this: that our taxes should be high enough to protect the stability of our currency, and no higher.... Now it follows from this principle that our tax rates can and should be lowered to the point where the federal budget will be balanced at what we would consider a satisfactory level of high employment.*

This principle is also one adopted in this textbook, but with one caveat. Ruml was addressing the situation in which the external sector balance could be ignored (which was not unreasonable in the case of the US in the early post-war period). In today's world, in which some countries have very high current account surpluses and others have high current account deficits, the principle must be modified.

We would restate it as follows: **tax rates should be set so that the government's fiscal outcome (whether in deficit, balanced, or in surplus) is consistent with full employment.**

Nations that will typically have a current account deficit at full employment (such as Australia, USA, UK) will normally have a fiscal deficit at full employment (equal to the sum of the current account deficit and the domestic private sector surplus).

Countries like Japan (with a current account surplus at full employment) will have a smaller fiscal deficit at full employment (equal to the domestic private sector surplus less the current account surplus). Countries with larger current account surpluses at full employment, such as Norway, will typically have a fiscal surplus at full employment, so as not to promote inflation.

20.6 Currency Sovereignty and Policy Independence

Currency-issuing nations such as the US, the UK, Australia, Japan, Turkey, and Argentina after it abandoned the currency board, or Italy before it joined the Eurozone, created **a currency for domestic use**. The government itself (including the treasury and the central bank) issues, spends, and lends the monetary base, including coins, notes and bank reserves as its liability.

These governments do not promise to convert their currency to any other currency, nor to gold or any other commodity, at any fixed exchange rate. The flexible exchange rate is a key to maintaining fiscal and monetary policy independence – what we will call sovereignty, although government sovereignty certainly has other dimensions as well.

By contrast, as we noted earlier, if a country pegs its exchange rate, it must hold sufficient foreign currency reserves to maintain the peg, which means that it must subsume domestic policy independence to the overriding necessity of accumulating reserves. It thus surrenders monetary sovereignty and hence domestic policy independence in the name of external balance. This is why **a floating exchange rate is a necessary component of policy independence.**

But there is more to it than a flexible exchange rate. The sovereign government spends (buys goods, services, assets, or makes transfer payments) by issuing a treasury cheque, or increasingly, by simply crediting a private bank deposit. In either case, however, credit balances (in the currency of issue) are created when the central bank credits the reserve account of the receiving bank.

Analogously, when the government receives tax payments, simultaneously the taxpayer's bank deposit is debited, and their bank's reserves at the central bank are reduced.

While it is usually supposed that the operation is reversed, with a government needing to first receive tax revenue, and then spend that revenue, this sequence is not necessary for any sovereign government in the absence of any voluntary, self-imposed constraints.

If a government spends by crediting a bank account (issuing its own IOU or currency of issue) and taxes by debiting a bank account (and eliminating its IOU or currency of issue), then it is not spending tax revenue. With a floating exchange rate and a domestic currency, the sovereign government's ability to make payments is not revenue constrained precisely because it spends by emitting IOUs.

Note that the sale of its own debt by a sovereign government should not be thought of as a borrowing operation, even though it is frequently described as such. As discussed in the previous sections, the operational effect of government bond sales (whether by the treasury in the new issue market, or by the central bank in open market operations) is to drain any excess reserves created (mostly) by treasury deficit spending. If the bond sales were not undertaken to drain excess reserves, the overnight rate would fall to zero (or to the interest rate paid on reserves by the central bank).

The treasury and the central bank work together to ensure that the overnight interest rate target (set by monetary policy) is maintained. They do this through sales or purchases of government bonds to drain or add reserves as necessary to allow the monetary authorities to manage liquidity (reserves) and balance supply and demand for reserves at the desired target interest rate.

When a household or non-sovereign government borrows, it issues an IOU and obtains a bank IOU that it needs in order to spend. The sovereign government, on the other hand, has no need to obtain a deposit at a bank before it spends its own currency. It can spend by issuing currency directly, or by crediting a private bank account. It sells a security, not to finance its expenditures, but to reduce the outstanding stock of currency of issue in the non-government sector, offering to substitute one of its interest paying liabilities (the bond) for a non- or low-interest paying liability (the currency usually in the form of central bank reserves) that is then debited from bank accounts.

This is really an interest rate management operation, reducing bank reserves in order to eliminate (non-interest earning) excess reserves that would otherwise place downward pressure on overnight interest rates. As such, **bond sales are really a part of monetary policy, and not a required part of fiscal policy**.

The final point to be made regarding such operations by a sovereign government is that the interest rate paid on government bonds is not subject to normal market forces. The sovereign government could always choose to leave excess reserves in the banking system, in which case the overnight rate would fall toward zero (or the support rate, the interest rate paid by the central bank on reserves).

When the overnight rate is zero, the treasury can always offer to sell short-term bonds that pay a few basis points above zero and will find willing buyers because such bonds offer a better return than the alternative (zero). This drives home the point that a sovereign government with a floating currency can issue bonds at any rate it desires, normally a few basis points above the overnight interest rate target it has set.

Moreover, the central bank can set whatever yield it chooses on outstanding government debt by offering to buy unlimited quantities in the secondary bond markets. We will return to this in Chapter 23.

There may be economic or political reasons for keeping the overnight rate above zero (which means the interest rate paid on bonds will also be above zero). But it is not correct to argue that the size of a sovereign government deficit affects the interest rate paid on the bonds that it issues.

Not understanding this, treasuries sometimes try to play the yield curve, issuing longer maturities when interest rates on them are low, or reversing course and issuing short maturities when the yield curve is steep.

While it could be true that market forces of supply and demand enter into maturity spreads, if treasuries understood that the purpose of bond sales is to drain excess reserves so that the central bank can hit its overnight interest rate target, they would not issue long maturity debt at all. Indeed, paying interest on reserves is an adequate substitute for treasury debt issue, as the overnight rate cannot fall below the interest rate on reserves.

Conclusion

This chapter provided an introduction to monetary and fiscal policy operations where governments issue their own sovereign currency. The central bank runs the payments system, ensuring banks have the reserves they need for cheques to clear. The treasury is the fiscal arm of the government, making payments for government that are authorised through the budgeting process. In modern countries, payments by and to the treasury go through the central bank, which means that the treasury and central bank must coordinate activities. While it is often claimed that central banks (such as the Fed) are, and must be, "independent" of the treasury, this independence is quite limited in practice, because of the necessity of ensuring treasury checks always clear. Central bank independence is largely limited to setting the overnight interest rate target; hitting that target also constrains independence of the central bank. Since government spending and taxes both impact bank reserves, the central bank normally offsets undesired impacts – requiring further coordination of central bank and treasury operations.

This chapter reinforced the conclusions from earlier chapters: **the sovereign government cannot run out of money.** In spite of operational rules that might be imposed on the treasury and central bank, procedures have been adopted to ensure that treasury can make all payments as they come due. Treasury cheques never 'bounce' due to insufficient funds. For these reasons, it is often useful to consolidate the balance sheets of the treasury and central bank for analysis of the impact of government spending and taxing (and, thus, government deficits) on the non-government sectors. Finally we discussed why floating currencies preserve more space for policy independence. We will take up many of these themes in more detail later.

References

Ball, L. and Sheridan, N. (2003) "Does Inflation Targeting Matter?" NBER Working Paper No. 9577, March.

FOMC (Federal Open Market Committee) (2016) "Statement on Longer-Run Goals and Monetary Policy Strategy" available at: http://www.federalreserve.gov/monetarypolicy/files/FOMC_LongerRunGoals_20160126.pdf, accessed 19 February 2016.

Lavoie, M. (2013). "The Monetary and Fiscal Nexus of Neo-Chartalism: A Friendly Critique." *Journal of Economic Issues*, 47(1), 1–31.

Ruml, B. (1946a) "Taxes for Revenue are Obsolete", *American Affairs*, Winter Number, VIII(1). Available at: http://www.constitution.org/tax/us-ic/cmt/ruml_obsolete.pdf, accessed 2 May 2017.

Ruml, B. (1946b) "Tax Policies for Prosperity", *The Journal of Finance*, 1(1), 81–90.

Ruml, B. (1950) "The New Economic Insight", *American Affairs*, XII(2), 90–3.

Tymoigne, E. and Wray, L.R. (2013) "Modern Monetary Theory 101: A Reply to Critics." Levy Economics Institute of Bard College, Working Paper no. 778, November.

Endnotes

1. The 'statutory mandate' was imposed by Congress in the Federal Reserve Act, and amended in 1977 to read: "maintain long run growth of the monetary and credit aggregates commensurate with the economy's long run potential to increase production, so as to promote effectively the goals of maximum employment, stable prices and moderate long-term interest rates." The first two of these are referred to as the 'dual mandate'; the third is considered to be more of a tool to achieve the mandate.
2. The 'discount window' is the central bank's lending facility. The term comes from the historical practice of submitting eligible short-term financial assets at the central bank's 'window' in exchange for reserves. If the 'face value' of the asset is worth $100, payable in 90 days, the central bank might provide $98 of reserves, the 'discount' of $2 representing interest charged on the loan for the 90-day period. Today central banks lend against eligible collateral at the 'discount rate'.
3. We outline these practices in detail in Chapter 23. Note that if a central bank pays interest on reserves equal to the target rate, then excess reserves will not push the overnight rate below the rate paid by the central bank. Normally the central bank will quickly respond by removing excess reserves. However, after the Global Financial Crisis some central banks adopted a practice of leaving huge quantities of excess reserves in the banking system (keeping overnight rates at or even slightly below the floor of the target range) in a practice called 'quantitative easing'. We will discuss that in Chapter 23.
4. Here we include holdings of bank deposits and notes and coins.
5. Within the foreign sector there are nations that run current account deficits and others that run current account surpluses. The notable surplus countries are Germany, China, Japan, and other East Asian countries.
6. As noted earlier, following the GFC, central banks engaged in an extraordinary policy of 'quantitative easing' that maintained massive excess reserves in the banking system on the (largely misguided) belief this would stimulate private sector lending and spending.

7. The analysis can be further developed by assuming that the private sector wishes to hold an extra ten dollars of banknotes, which reduces their bank deposits and the additional reserves that the banks choose (need) to hold.
8. Note that in some textbooks, the monetary base is referred to as high powered money. We prefer not to use that terminology because it is intrinsically related to the flawed concept of the money multiplier which asserts that monetary base is multiplied into a large money supply through private bank credit creation (further discussion about this is found in Chapter 10).
9. The term derives from Latin and French, meaning 'to return'. What 'returns'? The government's own money. Of course it cannot 'return' unless it has already been spent, reflecting the logic that currency must be spent before taxes can be paid. Logically then, the taxes cannot be a source of 'revenue' in the usual sense of the meaning, as a precondition to spending. While true of households and firms, it is not true of the sovereign issuer of the currency.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER 20 APPENDIX: ADVANCED MATERIAL

Monetary policy in the open economy, causes and consequences of capital flows

Here we extend the analysis of central bank operations to consider the open economy. In much of our previous discussion of monetary policy, we assumed that the non-government sector was equivalent to the private domestic sector. However, we must take account of international payments and possible effects on bank reserves. We will conclude that while this complicates the exposition, it does not significantly change the results.

Sterilisation of capital flows

The first issue explored is a practice called **sterilisation of capital flows**. The basic idea is that international payment flows can affect bank reserves. The question then becomes: Should the central bank offset these impacts? In other words, should it sterilise those impacts? If these flows increase reserves, should the central bank take the extra reserves out of the banks? If the flows reduce reserves, should the central bank restore reserves?

In the orthodox literature this is presented as a choice. The fear is that if the flows increase reserves, and if the central bank does not take out those extra reserves, then the banks will increase domestic lending, creating money that could cause inflation. Hence the orthodox recommendation is that the flows should be sterilised.

Notwithstanding the observation that banks do not loan out reserves to customers, our response is that **sterilisation is more or less automatic.** This is because a central bank that targets overnight interest rates must accommodate reserve demand or it will miss its target. Let us turn to the detailed argument.

If international payments flows (or domestic fiscal actions) leave banks with excess reserves, the central bank has no choice but to drain the excess unless it is willing to allow the overnight rate to fall towards zero, or it pays an interest rate on reserves equal to the target interbank rate. As we have learned, draining reserves is accomplished through open market bond sales, unwinding discount window lending, or sales of foreign currency reserves.

On the other hand, if international payment flows or domestic fiscal actions leave banks with insufficient reserves, overnight rates would rise above target, which triggers the opposite interventions.

For this reason, sterilisation is not a discretionary operation. For example, Japan currently (2018) runs a large trade surplus with the US. Japanese importers want to convert their dollar-denominated receipts to Yen, an operation that is facilitated by the central bank when it buys dollars and creates Yen reserves.

If this leads to excess reserves in the Japanese banking system, the central bank then drains the excess through, for example, a sale of Japanese government debt. It cannot choose, however, to leave excess reserves in the banking system unless it is prepared either to see the overnight interest rate fall toward zero or to offer a competitive interest return to the banks on their excess reserve holdings. In the absence of offering such a support rate, any sterilisation of Yen reserves is automatic, a result of the interest rate targeting procedure.

Government deficit finance and capital inflows

It is sometimes claimed that a government's deficit spending, as well as a nation's external position, are constrained by the portfolio preferences of savers. For example, many believe that government faces a fiscal constraint, according to which its spending must be financed by a combination of tax revenues, bond sales (borrowing), or money creation. The form that financing of fiscal deficits takes is thus supposed to depend on the portfolio preferences of savers.

It is claimed that once households and firms have accepted all the new money desired, government must sell bonds, and the interest rate required to get the public to hold the bonds will be determined by their preferences.

This supposedly applies even more forcefully to external constraints on national government fiscal deficits. For example, it is often erroneously claimed that the foreign sector (particularly China) is financing the US fiscal deficit by lending US dollars to the government.

It is feared that once the ROW (Rest of World) has all the US government bonds that it desires, the US government won't be able to finance its deficit except at rising interest rates.

Finally, it is argued that the ROW might even turn against the US dollar, refusing to hold dollars or US government debt, resulting in a financing crisis for the US and its government. Similar arguments have been made about other nations running external deficits.

This thinking reflects several different types of confusion. First, it conflates saving (consumption less than income) with portfolio allocation decisions (in what form wealth is held). Second, it inappropriately equates the position of the issuer of the currency (the sovereign government) with the user of a currency (domestic households and firms, plus foreigners). Third, it applies an analysis that might be appropriate for a nation on a fixed exchange rate regime to a nation operating with a floating currency.

A sovereign government running a floating rate regime spends by crediting bank accounts, so the government fiscal constraint is nothing more than an *ex post* accounting identity (see Chapter 21). If a deficit results, government drains any excess reserves through bond sales as part of its interest rate targeting procedure. Again, a nation that pays interest on reserves never needs to sell bonds because the interest earning reserves serve the same purpose as interest paying bonds.

The non-government sector makes its portfolio preferences apparent as excess reserves drive the overnight interest rate below the target rate and will accept government bonds until all undesired reserves are drained. The demand for reserves is highly interest inelastic, but even if it were not, government can set the overnight rate at any positive level desired simply by ensuring that the banking system has no more reserves than it wants.

It is important to understand that the actual fiscal outcome in any period is not under the control of the government of the day. It is an endogenous outcome determined by a combination of the discretionary policy choices taken by the government and the spending and saving behaviour of the non-government sector.

Whether the *ex post* fiscal identity will record a deficit after government increases its spending depends largely on the reaction of the other sectors. In other words, the government can decide how much it will increase spending and after the fact we will observe some combination of increased tax revenue, increased bonds held by the non-government sector, and increased base money holdings (reserves held by banks and cash held by the non-banking private sector).

The degree to which taxes rise will depend on the responsiveness of tax revenue to rising aggregate spending and income; the additions of bonds and the monetary base to non-government portfolios will equal (by identity) the fiscal deficit, and the split between the two will depend on preferences for interest earning assets, given the overnight interest rate set by central bank policy.

The saving propensities of both the private domestic sector and external sector go into determining the financial balances of all three sectors: the domestic private sector, the foreign sector, and the government sector.

Higher private domestic sector saving overall represents a leakage out of domestic income that is matched by some combination of a bigger government deficit and a smaller current account deficit. Higher ROW saving is matched by a combination of a larger government deficit and a greater current account deficit (holding the domestic private sector's balance constant).

We cannot observe saving or import propensities and our three-sector balances identity cannot tell us the complex causalities that lie behind the resulting balances. However, we should understand that **the fiscal outcome for a currency-issuing government is largely a residual**, rising when private domestic and foreign demand shrinks and falling when demand is rising. By the same token, a nation's current account deficit is largely a function of the ROW desire to spend.

Unfortunately, most analysts incorrectly interpret the causal forces involved, adopting a loanable funds approach, according to which saving 'finances' investment, fiscal deficits, and current account deficits. Actually, the causation is (mostly) the reverse: it is the investment spending, the government spending, and the export spending that together create the domestic saving of the private sector and the foreign saving in the form of the currency of issue.

Higher spending brings forth its own additional saving as a result of income growth.

A moment's reflection about bank balance sheets will confirm that this must be true. A saver cannot simply ask their bank to credit their savings account with more dollars, but an investor can approach a bank for a loan, in which case the investor's deposit account is credited and this transaction is offset on the bank's balance sheet by the loan, which is the bank's asset. When the investor purchases plant and equipment, that deposit account is drawn down and a saver's account is credited. Investment *creates* savings.

Similarly, a foreigner cannot save more dollars (local currency) until a local importer has purchased foreign output (or purchased foreign assets, including direct investment). Again, it is the importer's willingness to take out a loan to finance this purchase that results in a new dollar credit to the account of the foreign saver. Hence, the notion that a nation is borrowing its local currency (for example, US dollars in the case of America) from abroad (for example, from China in the case of the US) to finance government and trade deficits, is erroneous. Rather, it is more revealing to think of the nation's fiscal deficit and the current account deficit as financing the ROW saving in that currency (which is identically equal to the current account deficit).

The decision to save is a decision to 'not spend'. For example, when the Japanese private domestic sector, taken as a whole, produces more than its government and non-government sectors wish to purchase, it can save in financial form, but only if it can find external buyers so that it can export. Otherwise, saving takes the form of undesired inventory accumulation, which would then probably depress future production, employment, and income.

Let us assume Japan sells the excess production to Americans, in which case the savings are initially in US dollars. Portfolio decisions then come into play when Japanese savers decide how to hold the savings. Most of the US dollars will be exchanged for yen and used to purchase yen assets (financial and real). The Bank of Japan will usually facilitate this process as domestic Japanese banks offer US dollar deposits (claims on US banks that will be converted to reserve deposits at the Fed), for yen reserves (claims on the Bank of Japan). As discussed above, if excess yen reserves result, these can be drained by the Bank of Japan if it desired to maintain a positive overnight interest rate. In reality, the Bank of Japan currently (2018) operates with a zero interest rate target and thus is happy to leave some excess reserves in the banking system.

In this situation, the portfolio decisions of foreigners (including importantly, those decisions of ROW central banks) place no direct pressure on the US overnight interest rate. However, they can affect the exchange rate of the US dollar. It is commonly believed that a nation that runs a trade deficit must eventually see its currency depreciate in foreign exchange markets, although it is well recognised that empirical studies have not been able to systematically link exchange rates to the usual set of variables thought to be important determinants of exchange rates, including the trade balance.

In any case, this is a separate issue from the concerns with interest rate setting by the central bank or 'financing' of external and fiscal deficits. A country with a sovereign currency on a floating exchange rate can set its policy interest rate at any level desired, and can run fiscal deficits at any level desired, without worrying about the impacts of foreign saving propensities or portfolio preferences with respect to 'financing'. The country might, if desired, adjust interest rates or fiscal policy in response to actual or supposed pressure on exchange rates. But that is again, a separate issue from 'financing'.

We conclude that **allowing for open economy impacts does not change our results** for the following reasons.

- Central banks will sterilise any impacts on banking system reserves in order to hit overnight interest rate targets. Alternatively, they can simply offer a support rate on excess reserves and leave them in the system.
- Foreign portfolio preferences can impact current account outcomes. If the ROW wants US and Australian dollar assets, it exports output to the US and Australia. In that case, the US and Australia might record current account deficits and the respective currencies will flow out to foreigners. On the other side of the coin, they will record capital account surpluses as dollars (US and Australian, respectively) flow in.

CHAPTER

21 FISCAL POLICY IN SOVEREIGN NATIONS

Chapter Outline

Learning Objectives

- Comprehend the competing perspectives about the conduct of fiscal policy (hawks, doves and owls).
- Understand the debates about fiscal policy as the alleged cause of crowding out and (hyper)inflation.

21.1 Introduction

In Chapter 20, we provided a detailed account as to how fiscal policy is conducted and the role of the central bank in interest rate (or liquidity) management through buying and selling treasury debt from or to the private banks.

In this chapter, we continue our analysis of fiscal policy in sovereign currency-issuing nations. We first contrast the MMT view that the implementation of fiscal policy should be based on Abba Lerner's **functional finance principles** with the thoughts of the orthodox economists who support the adoption of **sound finance** principles. We will see that the mainstream arguments about constraints on government spending follow from an inappropriate application of the **household budget constraint** to a sovereign government.

Our analysis leads to an important policy recommendation that government can and should design its programme of spending and taxation with the objective of achieving and maintaining full employment and price stability.

We then turn to key debates which arise from this central role for fiscal policy:

- **Crowding out** – we briefly explore the claim that expansionary fiscal policy 'crowds out' private sector spending via interest rate increases.
- **Inflation** – the great fear of mainstream macroeconomists is that the achievement of full employment will cause inflation which could deteriorate into hyperinflation. MMT theorists take a different view, and in Chapter 19 we discussed in detail a programme that can achieve full employment without leading to inflation: the Job Guarantee. Mainstream economists often use the examples of the Weimar Republic in the 1920s and

Zimbabwe in the late 1990s and early 2000s to justify their assertion regarding hyperinflation. We acknowledge that in the absence of appropriate oversight, a government can maintain an excessive rate of expenditure which leads to rising inflation. But we show that the two popular examples of hyperinflation – the Weimar Republic and Zimbabwe – were the result of increasing aggregate supply constraints rather than being driven by excessive fiscal deficits.

21.2 Functional Finance versus Sound Finance

The fiscal constraint and the views of deficit hawks, doves, and owls

Following the Global Financial Crisis (GFC), many developed economies experienced a rise in their fiscal deficit-to-GDP ratios due to the operation of automatic stabilisers following the collapse of non-governmental expenditure. Discretionary fiscal stimulus packages also increased the fiscal deficits.

These fiscal deficits translated into rising debt-to-GDP ratios. Since 2009 multinational agencies including the OECD and IMF have preached the principles of sound finance through the adoption of austerity measures. These are policies of cutting government expenditure and/or raising taxes that are required to reduce the fiscal deficit in the misguided belief that governments should balance their fiscal positions.

This represents the typical orthodox position on fiscal policy, with the **government fiscal constraint** viewed as an analogue to the household budget constraint derived from neoclassical microeconomic theory. The interpretation of the so-called constraint on government net spending is based on the premise that the government has three sources of finance for its spending, as shown on the right-hand side of the following identity:

$$(21.1) \qquad G + iB \equiv T + \Delta B + \Delta M_h$$

where G stands for government spending; iB are the interest payments on existing public debt; T is tax revenue, ΔB is new borrowing based on selling government bonds, and ΔM_h is new base money creation. This relationship was introduced in Chapter 20 where it linked the acquisition or loss of net financial assets of the domestic private sector to the government sectoral balance in a closed economy.

According to the mainstream interpretation of the **government fiscal constraint**, if the government runs a fiscal deficit (spends more than it receives in tax revenue), then it will have to borrow (ΔB) by selling public debt and/or create additional base money (ΔM_h). Thus, the government fiscal constraint is alleged to represent an *ex ante* constraint on government spending. In other words, if the fiscal outcome was known in advance, then Equation (21.1) would be a guide to how it could be financed.

Few economists argue that government must or even should continuously achieve fiscal balance, although there are occasionally some politicians and other extremists who want to legislate such a requirement.

We can distinguish three different perspectives as to the appropriate fiscal strategy: (a) deficit hawks; (b) deficit doves and (c) deficit owls. Category (c) was added by MMT economist Stephanie Kelton from UMKC in the USA.

Deficit hawks recommend that government strives to achieve fiscal balance or even surplus, even though most recognise that it is hard to exactly match revenues and expenditures over the course of a year. Hence deviations from fiscal balance will occur, but government should always respond to such imbalances. Thus, if a deficit occurs in one year, government should try to run a surplus in the following year to offset it by cutting spending and/or raising taxes.

Deficit doves believe government should aim to achieve fiscal balance over the course of an economic cycle, but should run deficits in recessions and offsetting surpluses in expansions. Hence, a government should be willing to use its fiscal capacity as a countercyclical policy tool to offset swings of private sector spending. For example, deficit doves argued for deficits to stimulate the slumping economies of the major Western nations during the GFC. In their view, the time to move towards fiscal balance would come only after a robust recovery had gotten underway and tax revenues had started to increase.

Deficit owls take an entirely different position, based on **functional finance** principles. For them, the fiscal outcome for a sovereign government is not a useful target for policymaking. It is not **functional** in the sense of providing policy guidance. Rather, policy ought to target important economic goals such as full employment, price stability, poverty alleviation, reduced income inequality, financial stability, environmental sustainability, and the overall standard of living.

Why is the deficit owl the only perspective that is consistent with MMT?

As we know from previous chapters, the sovereign government spends by issuing currency, which today is mostly through electronic entries on balance sheets. Taxes lead to debits of those entries. Logically, the spending must precede the taxing since accounts cannot be debited before they are credited.

MMT sees the government fiscal outcome as an *ex post* identity. At the end of a year, it will certainly be true that government spending over the year is equal to tax revenues plus net bonds issued plus net base money issued, as shown in Equation (21.1). In this sense, the equation becomes a simple accounting identity that must hold by definition, but it has no further merit.

Equally importantly, MMT does not see tax revenue (T), new base money creation (ΔM_H), and new borrowing based on selling government bonds (ΔB) as alternative methods of financing government spending. Instead, it sees them as different parts of the process of conducting fiscal policy, as described in Chapter 20, and illustrated by the numerical example.

Spending begins with crediting private bank accounts with central bank reserves. The payment of taxes leads to private bank accounts being debited. Then, if government spending is greater than taxes, there is a net crediting of reserve accounts at the central bank ($\Delta M_H > 0$).

Normally the reserves created will be greater than what banks need to hold, whether or not there are legal reserve requirements. Banks with excess reserves at the central bank will try to lend them in the interbank overnight lending market. However, when the overall banking system has excess reserves, there will be no demand for them. No individual bank can solve the problem of a system-wide excess of reserves. There must be a system-wide solution.

We understand from earlier chapters that an excess supply of reserves, in the absence of a support rate paid by the central bank, will cause the overnight interbank lending rate to fall. Once it has fallen below the target range, the central bank will respond by selling bonds (that is, via an open market sale). However, in normal periods central banks have a limited supply of government bonds; they can only sell bonds that they have previously bought.

So, in the presence of fiscal deficits, the central bank would need the treasury to create and sell more bonds in the primary issue market. Central banks and treasuries coordinate their operations to ensure that fiscal operations have minimal undesired impacts on banking system reserves. Hence bonds will be issued more or less in step with fiscal deficits in order to drain from the banking system excess reserves that would normally result from spending in excess of tax receipts.

At the end of the year, we would find that government spending less taxes will be equal to the change in base money, that is, the change to banking system reserves plus the change to private sector holding of cash, and in addition, the change to non-government holding of government bonds.

As noted, in normal times, the growth of banking system desired and/or required reserves is quite small. The growth of cash held by the non-government sector is likewise fairly small, and linked more or less closely to growth of national income. Thus, the deficit is usually approximately equal to the additional sales of government bonds (ΔB).

On the other hand, let us assume that a sovereign government spent by crediting banks with reserves but chose to leave excess reserves in the banking system (that is, it did not sell bonds). This could happen if, for example, government adopted a **zero interest rate target**; in that case excess reserves would drive the overnight interbank lending rate toward zero and the government would not need to do anything further. We would then see that $G + iB - T = \Delta M_H$. But the difference between this case and the more usual case discussed above, where $G - T$ is approximately ΔB, has nothing to do with the way government chose to 'finance' its spending.

In both cases, the government spent by crediting bank accounts. The different outcome is due to the choice between draining excess reserves or leaving excess reserves in the system. That decision depends on whether government wants to target a positive overnight interest rate, or a zero overnight interest rate.[1]

Most economists see the choice as a monetary policy decision, and not a fiscal policy decision. In the post-GFC period the level of bank reserves significantly increased in a number of developed economies, including Japan, the USA, and the UK. This reflected rising risk aversion among banks in these countries, and also the desire of their central banks to both maintain a low overnight rate and also flatten the yield curve (reduce longer-run interest rates) via **quantitative easing** (which we discuss in more detail in Chapter 23).

We conclude that the **government fiscal constraint** is neither a constraint nor does it present alternative ways of financing government spending. Rather it is an *ex post* accounting identity whose outcome is determined by decisions made by households, firms, financial institutions, the central bank, and even foreigner investors.

- Households, firms, and foreigner investors decide how much cash they want to hold. Banks (and the central bank through required reserve ratios in the USA) determine the level of reserves that they hold.
- The central bank decides whether the overnight rate target will be above zero.

All of those decisions go into determining the split between ΔB and ΔM_h.

That is not an *ex ante* decision of treasury to either **borrow** or **print money**. Indeed, treasury cannot decide *ex ante* what the fiscal outcome will be (fiscal balance, deficit, or surplus) since that depends on tax revenue generated over the course of the upcoming year, plus any unplanned spending linked to unforeseen events and the impact of automatic stabilisers.

Equation (21.1) is thus useless for planning purposes, and has no explanatory power *ex post* in terms of explaining the composition of its right-hand side.

Functional finance

In the 1940s, American economist Abba Lerner wrote two important articles which still resonate today. One proclaimed that "money is a creature of the state" (Lerner, 1947: 313). Obviously, that is also the position of MMT as outlined in this textbook: the state chooses a money of account, imposes liabilities in that unit, and issues currency denominated in the same unit that is accepted in payment of taxes. All of this was understood by Lerner.

In his article on functional finance, he calls it the new fiscal theory. He says that like any new theory it seems extremely simple and it is that simplicity that makes people suspicious. Lerner wrote (1943: 39): "The central idea is that government fiscal policy, its spending and taxing, its borrowing and repayment of loans, its issue of new money, and its withdrawal of money, shall all be undertaken with an eye only to the results of these actions on the economy and not to any established traditional doctrine about what is sound or unsound."

He went on to outline two principles of functional finance:

1. "The first financial responsibility of the government (since nobody else can undertake that responsibility) is to keep the total rate of spending in the country on goods and services neither greater nor less than that rate which at the current prices would buy all the goods that it is possible to produce" (Lerner, 1943: 39). When spending is too high, the government is to reduce spending and raise taxes; when spending is too low, the government should increase spending and lower taxes.

2. "An interesting corollary is that taxing is never to be undertaken merely because the government needs to make money payments ... Taxation should therefore be imposed only when it is desirable that the taxpayers shall have less money to spend." (Lerner, 1943: 40)

If the government is not to use taxes to 'make money payments', then how are these to be made? According to Lerner, the government should not turn to borrowing for the purposes of spending because: "The second law of Functional Finance is that the government should borrow money only if it is desirable that the public should have less money and more government bonds" (Lerner, 1943: 40).

In other words, the purpose of taxes and bonds is not to finance spending, as each serves a different purpose (taxes remove excessive private income while bonds offer an interest earning alternative to money). Instead, the government should meet its needs "by printing new money" (Lerner, 1943: 41) whenever the first and second principles of functional finance dictate that neither taxes nor bond sales are required. That is, as discussed above, the choice over whether to leave base money (mostly reserves) in the system, or to drain it through bond sales, depends on what we normally call a monetary policy decision (interest rate policy). In summary, Lerner argued (1943: 41):

> *Functional Finance rejects completely the traditional doctrines of 'sound finance' and the principle of trying to balance the budget over a solar year or any other arbitrary period. In their place it prescribes: first, the adjustment of total spending (by everybody in the economy, including the government) in order to eliminate both unemployment and inflation, using government spending when total spending is too low and taxation when total spending is too high; second, the adjustment of public holdings of money and of government bonds, by government borrowing or debt repayment, in order to achieve the rate of interest which results in the most desirable level of investment; and third, the printing, hoarding or destruction of money as needed for carrying out the first two parts of the program.*

He concluded that functional finance "is applicable to any society in which money is used as an important element in the economic mechanism" (Lerner, 1943: 50).

In this textbook, we want to narrow the application of functional finance somewhat to the sovereign government, which is consistent with the views Lerner advanced in his other great article, published in 1947. In that piece, Lerner insisted that: "[W]hatever may have been the history of gold, at the present time, in a normally well-working economy, money is a creature of the state. Its general acceptability, which is its all-important attribute, stands or falls by its acceptability by the state" (1947: 313).

Just how does the state demonstrate acceptability? For Lerner:

> *The modern state can make anything it chooses generally acceptable as money ... It is true that a simple declaration that such and such is money will not do, even if backed by the most convincing constitutional evidence of the state's absolute sovereignty. But if the state is willing to accept the proposed money in payment of taxes and other obligations to itself the trick is done. Everyone who has obligations to the state will be willing to accept the pieces of paper with which he can settle the obligations, and all other people will be willing to accept these pieces of paper because they know that the taxpayers, etc., will accept them in turn. (Lerner, 1947: 313)*

This seems to be a clear exposition of what we now call the MMT, **taxes drive sovereign currency** view. Even if it has not always been the case, it surely is now true and obvious that the state chooses and enforces the money of account when it denominates the tax liability in that money of account, and defines what will serve as 'money' when it decides what will be accepted at tax offices. The 'money' is widely accepted not because of sovereignty alone, not (simply) because of legal tender laws, and not because it might have (or have had) gold backing, but because the state has the power to impose and enforce tax liabilities and because it has the right to define 'that which is necessary to pay taxes'.

There can be no doubt that all modern states do have these rights. As Lerner said "Cigarette money and foreign money can come into wide use only when the normal money and the economy in general is in a state of chaos" (Lerner, 1947: 313).

One might only add that when the state is in crisis, loses legitimacy, and loses its power to impose and enforce tax liabilities, 'normal money' will be in a 'state of chaos', leading, for example, to the use of foreign currencies in private domestic transactions. In all other cases, it is state money which is used, and state money that the state accepts in payment of taxes.

21.3 Fiscal Policy Debates: Crowding Out and (Hyper)Inflation

Crowding out?

Many pundits wrongly believe that when national governments issue treasury debt, it can **crowd out** private borrowing and hence spending.[2] The basic premise is that there is a limited supply of private sector saving for which government **borrowing** and private sector borrowing compete. If government tries to borrow more by issuing and selling more bonds, then the competition for finance would push up interest rates. Some private firms would decide not to borrow at the higher rates and hence investment would be lower. Also, durable consumption expenditure by households will decline because some of it is financed by borrowing.

However, that model is incorrect. Government fiscal deficits generate non-government surpluses (flows) that accumulate to the non-government sector's net acquisition of financial assets (a stock), as we have learned in this and earlier chapters. Since there are more savings (as a result of higher income levels arising from the deficits) and greater financial wealth, it is not true that government is competing with private sector borrowers for a limited supply of saving (a flow) to place government bonds into wealth portfolios that are fixed in size. Both savings and portfolios expand as government deficits grow.

Additionally, as we learned in Chapter 10, commercial bank lending is not reserve constrained. Loans create deposits and any creditworthy borrower can access funds for investment expenditure, irrespective of the conduct of fiscal policy by the government. It is also the case that investment spending is typically weak when the economy is not growing robustly (see Chapter 25) so there tends to be a 'crowding in' impact arising from fiscal deficits in downturns, as the deficits attenuate recessionary pressures and encourage the recovery of investment.

Further, as we have seen, the central bank department of government sets the overnight interest rate. It can keep that low no matter how big the government deficits are. For example, this is why Japan has kept a near-zero overnight interest rate (and also very low rates on its long maturity treasury debt) ever since its economy collapsed in the early 1990s, leading to the biggest fiscal deficits in the developed world. (See the discussion of Japan's deficit, debt and interest rate outcomes in Chapter 2.) Similarly, the US central bank (the 'Fed') kept US interest rates low after the GFC, in spite of fiscal deficits that rose to ten per cent of GDP. Finally, all through the Second World War the Fed also kept rates near zero as fiscal deficits reached 25 per cent of GDP.

What all this means is that there is no reason to expect fiscal deficits to push up interest rates because interest rates (at least at the short end of the maturity structure) are policy determined.

For these and other reasons, the **crowding out** argument against fiscal deficits is not based on a coherent understanding of operational realities, or of the empirical data. The notion that government bonds compete for a fixed supply of saving must be rejected. It would be closer to the truth to assume that the demand is perfectly elastic for sovereign government securities issued when government runs deficits.

Indeed, primary dealers finance their purchases of bonds at auction in the repo market, mostly using treasury bonds as collateral, while the newly issued bonds will likely serve as collateral for further credit creation in financial markets.[3] Far from crowding out, bonds can enable more private credit creation than would occur in their absence.

Voluntary constraints

In Chapter 20 we made the simplifying assumption that the treasury could sell debt to the central bank on the primary market, to ensure that it had sufficient balances in its account at the central bank to undertake its desired level of expenditure. There we stripped away the veil of ideology to expose the capacity of the consolidated treasury/central bank in a modern monetary economy to achieve its fiscal policy objectives without being subject to voluntary constraints.

On the other hand, an assessment of the way fiscal policy is actually conducted in say the UK, USA or Australia requires a careful examination of the particular institutional practices which operate in those countries. The key question is whether the imposition of voluntary constraints imposes meaningful limitations on the treasury of a government that wants to pursue a full employment agenda.

In all three countries, the central bank does not make major purchases of public debt in the primary market. In other words, the central bank does not normally buy the treasury securities directly from the treasury, but rather buys them in 'secondary' markets from banks and other bondholders.

Debt management is practised through the regular issue and sale by auction to the non-government sector of short- and long-term public debt which is designed, in aggregate, to ensure the amount of debt issues matches the projected fiscal deficit, namely $G - T + iB$.

Note that we use the term 'fiscal deficit' whereas in the literature you will find that the phrase 'borrowing requirement' is typically used, although the 'borrowing' is required only due to existing voluntary operating procedures. Remember that government must spend (or lend) its currency before it can receive it back either in payment of taxes or in purchases of its debt. It does not really need to 'borrow' in its own currency, as it can always spend by crediting bank accounts. Further, as we discussed above, we cannot even know whether government will run a deficit until the end of the period. However, as we explain, under current procedures in many countries governments may choose to sell bonds before spending.

Sales of debt typically occur through an auction system, so that the quantities of short- and long-term debt are determined *ex ante* (in line with the anticipated fiscal deficit), and the market determines the interest rates on these assets.

An important question is whether this market exposure leads to high interest rates being paid, or in extreme circumstances, to the market's outright refusal to purchase the government debt on offer, which would lead to a shortfall of funds available to government under these voluntary arrangements.

Government bonds are riskless assets denominated in the domestic currency and hence constitute an important component of private portfolios. Non-government sector portfolios use the risk-free government debt as a benchmark. Consequently, the issue of public debt is typically oversubscribed, so interest rates associated with different maturities are relatively low, but higher than the prevailing target interbank rate.[4]

Central banks also have the capacity to manipulate longer-term yields, rather than just setting their target (overnight) interbank rates. They have the financial capacity to buy unlimited quantities of government debt in secondary markets, which pushes up prices and drives down yields. In turn, this influences the yields achieved in the primary market auctions.

Table 21.1 explores a sequence of stylised transactions, which are generally applicable to most nations, whereby the treasury department issues debt to private banks. When private banks buy treasury bills and bonds, the central bank debits their reserves and credits the treasury's deposit account at the central bank. In some nations (such as the US), the treasury holds deposits at private banks and then transfers them to the central bank prior to spending. We abstract from those specific arrangements in what follows.

We start with the (unrealistic) assumption that the central bank and the private banks have zero assets and liabilities. This assumption is only made so that the following sequence of transactions is as easy to understand as possible and doesn't alter the veracity of the analysis. We shall also assume that the prevailing voluntary institutional arrangements require the treasury to sell bonds to private banks before it can spend.

Let us say that the government issues $100 worth of treasury bonds to the private banks in a primary auction (Table 21.1, Stage 1). When a sale takes place, the central bank debits the reserves of the private banks and credits the treasury's deposit account with the central bank. The central bank records this transaction by recording a 'liability' of $100 as a treasury deposit and records a negative $100 amount against bank reserves. The banks which purchased the bonds record the $100-worth of bonds they now hold as assets and note that their reserves at the central bank are now deficient by $100 as a result of the purchase. At this stage, the banks are collectively, suffering from a shortage of reserves.

Where do they get these reserves from? The shortage can be addressed by the central bank purchasing the $100 of treasury bonds from the private banks on the secondary market (Stage 2). The banks no longer hold the $100-worth of bonds, but their reserve position is now balanced (zero). At this stage the government has yet to increase its spending and retains its deposit of $100 in its central bank account.

Stage 3 in Table 21.1 shows the balance sheet changes that follow from the government spending $100. We observe that the central bank now reduces the treasury deposits, which fall back to zero, and credits the reserve accounts of the banks by $100. The private banks record the increase in their bank reserves ($100) and the non-government, non-bank sector enjoys increased deposits of $100 as a result of the government spending on goods and services. Note that we have assumed that no tax receipts have occurred, so the government has incurred a deficit equal to $100.

As a consequence of the fiscal deficit spending, the banking system now has excess reserves of $100 and to avoid competition in the interbank market between individual banks trying to shed the excess, the

Table 21.1 The modified MMT view of government deficit spending

Stage	Central bank				Private banks			
	Assets		Liabilities		Assets		Liabilities	
1			Treasury deposits	100	Treasury bonds	100		
			Bank reserves	−100	Bank reserves	−100		
2	Treasury bonds	100	Treasury deposits	100	Treasury bonds	0		
			Bank reserves	0	Bank reserves	0		
3	Treasury bonds	100	Treasury deposits	0	Treasury bonds	0	Non-government sector deposits	100
			Bank reserves	100	Bank reserves	100		
4	Treasury bonds	10	Treasury deposits	0	Treasury bonds	90	Non-government sector deposits	100
			Bank reserves	10	Bank reserves	10		

Source: Adapted from Lavoie (2013: Table 3) with permission from Taylor & Francis Ltd. Column 1 has been added for clarity.

central bank sells $90 worth of treasury bonds to the private banks (who we assume desire to hold reserve balances of $10) (Stage 4).

The sequence of transactions allows the central bank to maintain control over its monetary policy interest rate target (assuming it to be non-zero) and avoid the need to pay interest on excess reserves. But there would be no substantive difference between the central bank mopping up the excess reserves in Stage 3 with open market operations (Stage 4) or choosing instead to leave the excess reserves in the system and reward them with a competitive yield. Either way the banks maintain an interest bearing asset of the same amount.

In Chapter 20 (Table 20.1) we analysed the impact of the treasury selling debt directly to its central bank, receiving a credit to its deposit account that it can spend down to advance its socio-economic programme. In Table 21.1, we modified that sequence by first selling the bonds to the private banks.

However, a comparison of the final balance sheet results of the two sequences (Table 20.1 compared to Table 21.1) shows that they are the same, despite the initial need to sell treasury bills to the non-government sector in the second example. Here the central bank buys the bonds in secondary markets from the private banks rather than buying them directly from the treasury in the primary market.

Inflation and sovereign fiscal policy

REMINDER BOX

Competing theories of inflation have underpinned the major macroeconomic policy debates over the last 50 years. In Chapter 17, the Quantity Theory of Money was outlined. In Chapter 18, the analysis commenced with the Keynesian Phillips curve trade-off and then introduced the expectations augmented Phillips curve, developed by Milton Friedman and Edmund Phelps. The chapter concluded with the hysteresis-based theory of inflation, which restored a trade-off between inflation and labour underutilisation.

In Chapter 19, we argued that there is a fundamental policy choice between the adoption of an unemployment buffer stock to sanction the inflationary process and the adoption of a fixed money wage, employment buffer stock which is the basis of the **Job Guarantee** proposal. We developed the argument that full employment is the primary policy goal, and after consideration of other strategies which were designed to achieve full employment, concluded that a Job Guarantee should be adopted.

We now focus again on the hostility of many economists to the active, discretionary use of fiscal policy and specifically the MMT view that there are no financing constraints on the use of fiscal policy, given that through keystrokes, treasury spends a fiat currency which by definition is not backed by any valuable commodity.

Many argue that it is the adoption of **fiat money** that causes inflation. If only the nation's money were tied to something with real value (like gold), that would allow money to retain its value so that prices would not rise.

On the other hand, MMT argues that money's value is not, and never has been, determined by a commodity like gold. Rather, money is the unit of account in which debts and credits are denominated. We can think of money as entries on balance sheets. Critics of MMT react in horror at such a suggestion, because they believe that MMT's claim that government spends through **keystrokes** is a recipe for inflation, if not hyperinflation. The following discussion should allay those fears.

REMINDER BOX

It is important to recall from Chapter 17 that **inflation is defined as a persistent rise in prices at the aggregate level**. The world experienced a significant inflation event in the early 1970s following the **energy shock** which was set off by geopolitical forces. Oil prices quadrupled over a short period, and most prices quickly rose because energy is used to produce and/or transport most goods and services.

In many countries of the developed world, there were strong trade unions who were seeking to restore their members' real wages, so an inflationary process commenced, which was typically met by contractionary macroeconomic policy and the emergence of stagflation (the simultaneous incidence of rising unemployment and rising inflation).

As Keynes argued, the economy needs some 'stickiness' of wages and prices in the money of account or people might abandon money. That is what can happen in a hyperinflation, when money's purchasing power is falling so quickly that people try to find a substitute that can maintain its value.

However, inflation in most developed economies has remained sufficiently low that the domestic currency has continued to be a useful money of account, and the domestic currency has been voluntarily held despite inflation occasionally rising to double digits. Economists have been hard pressed to find significant negative economic effects from inflation at rates of less than 40 per cent per annum. But clearly people do not like inflation when it rises to double digits, given the almost inevitable loss of real income, and policymakers usually react to double-digit inflation by adopting austerity programmes to reduce aggregate demand.

The question is whether austerity is the right policy tool. If an economy is operating beyond full employment, then by Lerner's first principle of functional finance, government needs to dampen demand by reducing spending or raising taxes. There have been instances in several countries over the past half-century in which demand probably became excessive, raising production beyond the full employment level. Major wars are the typical trigger for demand pull inflation. But in most developed countries, aggregate demand has been insufficient to generate inflationary pressure since the Second World War.

Many critics think that when advocates of MMT state that **government spends its currency into existence** and that logically it **spends first, then taxes** and that **for sovereign government, affordability is not a question**, they are making policy recommendations rather than describing a modern monetary economy. This is not correct because all of these statements apply to modern currency-issuing national governments.

All major economies went off the gold standard in the early 1970s. MMT's recommendation to use policy to pursue full employment simply follows the recommendations arising from Lerner's functional finance principles, namely to spend more if employment and incomes are too low, and to provide more reserves if banks are so short of reserves that it is driving up overnight interest rates and compromising monetary policy.

Insufficient spending and employment can be readily deduced when observing that there are jobless workers who cannot find jobs. These observations indicate that effective demand is too low, which means that government can either cut taxes or increase spending to raise demand. A sovereign government that issues its own currency can always afford to do this. The question is how best to stimulate demand.

Friedman famously used the metaphor of governments using helicopters to fly around dropping money into the economy to argue that this is how inflation is caused (Friedman, 1969). If government did inject money into the economy in this way, it would be a form of fiscal policy via **transfer spending** and involve the somewhat arbitrary distribution of welfare or social security payments (whoever finds a bag of money gets to spend it), rather than a monetary policy intervention (see Chapter 22).

Textbook Keynesian fiscal policy acts like this, and is often likened to a **pump priming** stimulus. It attempts to indiscriminately raise government spending and thus private income to encourage consumption.

In depressed economic conditions, this makes sense because it is likely that across all sectors of the economy there is sufficient slack. That allows for increased production and employment everywhere, with little pressure on prices or wages.

However, typically this is not the best form of expansionary fiscal policy to achieve full employment. The problem is that outside of the most severe depressions, normally some sectors have lots of slack, while others face relatively tight conditions.

This is particularly true across the labour force since the demand for educated and highly skilled workers is typically tighter than that for workers with less experience, education and training. Unemployment is never equally shared, and the most disadvantaged workers generally experience higher unemployment.

It is also the case that industrial locations are not evenly dispersed across space and so indiscriminate pump priming is not an appropriate approach to target regional disadvantage. Mitchell and Juniper (2007) introduced the term 'spatial Keynesianism' to describe an approach to fiscal policy that injects spatially focused stimulus to struggling regions with high unemployment.

Trying to **pump up** aggregate demand may well cause bottlenecks in some sectors that lead to workers negotiating for higher wages, which in turn, drives up prices in those sectors, even though there is substantial slack in other sectors. That means that inflation can be generated long before reaching full utilisation of plant and equipment and full employment of workers across most sectors of the economy.

As inflation rises, government can lose its nerve and abandon stimulus measures before full employment is achieved. In fact, given the lags in implementing fiscal policy, which were highlighted by Friedman, a commitment to the achievement of full employment by pump priming may be undermined by fiscal policy not operating as a countercyclical measure and destabilising the macro economy.

Friedman (1953) argued that discretionary fiscal policy is potentially destabilising because of long and variable time lags associated with: (a) recognition of the need for discretionary fiscal policy; (b) the design and implementation of an appropriate policy response; and (c) the economy responding to the policy measures that were adopted. In other words, by the time fiscal stimulus reaches the economy, the non-government spending cycle may have turned upwards and the stimulus would then become procyclical and risk being inflationary.

This is why Keynesians talk about governments planning ahead and having projects which are 'shovel ready' to bring into operation when the non-government spending cycle turns downwards.

Further, since pump priming entails the payment of market wages,[5] policymakers have limited scope for imposing a counter-inflation sanction while maintaining a commitment to full employment. This tends to lead to a **stop/go** pattern of policymaking, whereby government spends more to simulate demand but then slashes spending when inflation erupts.

What is needed instead is targeted policy that directs additional demand creation where it is most needed. This is not as hard as it might sound. Government does not need to keep tabs on every single sector of the economy to fine-tune its stimulus in order to help where it is most needed.

As we have discussed throughout the textbook, full employment is the most important policy goal for the economy, particularly if the UN Human Rights agenda is to be followed (see Chapter 1). In Chapter 19, we developed the **Job Guarantee** (JG) policy and noted that government can direct its spending where it is most needed – directly to the unemployed – by offering a job to anyone who needs one. Since government spending will automatically increase only when it is needed (that is, when the number of job seekers rises) and fall when it is not needed (when workers leave the programme for alternative work or leave the labour force), the policy operates countercyclically and only provides the amount of stimulus needed.

It is true that some factories and some retail outlets will still find themselves with excess or insufficient capacity. However, these will react to market signals by either cutting capacity or increasing it, as necessary.

Clearly there are always winners and losers in a market-based economy, but the private sector in total performs better when responding to market signals in a fully employed economy, which is achieved by putting into place the JG programme.

What about the argument that 'fiat money' is not backed by a real commodity such as gold, and is thus inherently inflationary? This fear reflects a belief that if government can buy anything that is for sale then it would try to buy everything.

However, there is ample evidence that at least in nominally democratic countries, governments exercise too much restraint in their spending and taxation measures. Usually the elected representatives in the legislative arms of government haggle over how much should be spent, and on what, and government spending decisions are subjected to widespread press scrutiny.

As we discussed in the previous section, governments impose all sorts of voluntary rules on themselves which constrain their fiscal latitude.

Further, the spending and tax legislation that ultimately emerges out of the political process does not determine what *ex post* spending and tax revenues will be since these depend on spending by the non-government sector, due to the operation of the **automatic stabilisers**. In a recession, more is spent on unemployment compensation as tax revenue declines; in an expansion, the reverse is true and revenue might exceed projections.

A government that decides to keep spending and raising the price it is willing to pay to purchase resources and output will undoubtedly cause high inflation. There is no substitute for good governance. An unaccountable government will not be constrained in its fiscal strategies either by the political process or by a gold standard. However, it is important to understand the cause of this inflationary process.

Monetarists are hostile to the creation of base money to finance deficits because they claim it is inflationary due to the Quantity Theory of Money (QTM). MMT advocates would first highlight institutional practice, namely that net treasury spending initially causes an equal increase in base money.

Second, they would challenge the theory of inflation based on QTM, and argue that if a fiscal deficit gives rise to demand pull inflation, then the *ex post* composition of $\Delta B + \Delta M_h$ in Equation (21.1) is irrelevant. Overall spending in the economy is the driver of the inflationary process, and not the *ex post* distribution of net financial assets created between bonds and base money.

Hyperinflation

Let us now look at instances of much higher inflation rates than those usually encountered in developed nations. These are inflation rates so high that they do harm to economies. We will see that extremely high inflation is unusual. Further, there appears to be no reason to believe that creeping inflation (that is common) will gradually rise to hyperinflationary rates (that are extremely rare).

Still, critics of MMT will cling to specific examples: either to the pre-Second World War Weimar Republic, or to the more recent experience in Zimbabwe, where it is claimed that government did spend without restraint, and in so doing destroyed the value of the currency.

Alternative explanations of hyperinflation

In recent years, several economists have advocated the use of **overt monetary financing** (OMF) to accompany fiscal policy. This policy requires the central bank to use its currency-issuing capacity to credit bank accounts on behalf of the treasury to facilitate public spending, without any matching public debt being issued to the non-government sector (Mitchell, 2015).

This policy approach is often referred to as **money printing**, but that is an erroneous descriptor because government spending is almost always accomplished via digital entries to banking systems rather than the physical creation of currency.

Many fear that if a government was to use OMF, then it would put the nation on a path to ruinous hyperinflation. The experiences of the Weimar Republic or Zimbabwean hyperinflations are often cited as reasons why fiscal deficits are dangerous.

These economic collapses were supposedly caused by governments that resorted to 'money printing' to finance burgeoning deficits, increasing the money supply at such a rapid pace that inflation accelerated to truly monumental rates.

In his classic 1956 study, Phillip Cagan defined hyperinflation as an inflation rate of 50 per cent or more per month. The most popular explanation of hyperinflation is the Monetarist quantity theory of money: the government engages in excessive money creation, causing prices to rise. However, as prices rise, the velocity of circulation increases, because no one wants to hold onto money for very long as its value falls rapidly. Wage increases are demanded daily, because tomorrow the same money wage will purchase significantly goods and services than today. Even though the money supply grows as rapidly as government can print notes, it never keeps up with rising prices. The higher that prices rise, the faster the velocity gets, until eventually workers demand hourly payment and run to the stores at lunchtime because by dinner time prices will be even higher.

Essentially, this was Cagan's explanation of the fact that a simple version of the quantity theory did not fit the data. If prices rise so much faster than the money supply, how can we conclude that the hyperinflation is caused by 'too much money chasing too few goods'? To fit the facts, the quantity theory was revised to state that in a high inflation environment, the old quantity theory presumption that velocity is stable (which is necessary to maintain a link between money and prices) no longer holds.

So armed with the revised quantity theory, it still could be claimed that high inflation and hyperinflation result from too much money, even though velocity rises when money growth lags behind the inflation rate. Monetarists claim that government controls the money supply, and hence hyperinflation must be due to government policy.

In addition, in hyperinflationary periods, the supply of government currency (paper notes) rises rapidly (sometimes with extra zeroes being added on the notes' values). Finally, government runs deficits when it finds its tax revenue cannot keep up with its spending, so it is alleged to frantically print money to make up the difference, and that adds to **too much money chasing too few goods** and generates accelerating inflation.

Thus, economists who criticise the MMT preference for OMF argue that most of the blame for hyperinflation falls on money printing by governments to finance deficits. Interestingly, these arguments were used in the Japanese policy response when its property market collapsed in the early 1990s. The Bank of Japan engaged in large scale **quantitative easing** (expanding bank reserves in exchange for government bonds) and the Ministry of Finance ran larger deficits; both policy interventions being designed to stimulate the economy. (We will consider quantitative easing in Chapter 23.)

The arguments were reprised during the GFC, in the US, UK and Japan, when large fiscal deficits (plus quantitative easing) led to a significant increase in the stocks of reserves held by the banks. It was argued that the massive build-up in bank reserves would lead to a lending frenzy which would generate accelerating inflation. History tells us that in both cases inflation did not accelerate.

Let's consider a MMT response to these explanations of hyperinflation to better understand why these historical predictions were inaccurate. We will make three points:

- As discussed above, when MMT says that government spends by **keystrokes**, this is a description, not a prescription. If critics were correct that government spending by **printing money** necessarily leads to high inflation or hyperinflation, then most developed nations would have at least high inflation, if not hyperinflation all the time because they all spend by keystrokes. Logically, all governments that issue their own currency have to spend it before they can collect it in taxes (or engage in bond sales), since no one else can create it. There is no alternative way for these governments to spend. Even if they promise to convert at a fixed exchange rate, they still spend by keystrokes. Yet there has been no evidence of hyperinflation or high inflation in developed economies over the last 20 years. This suggests that to claim a causal relationship between **printing money** and hyperinflation is highly problematic.
- Particular cases such as the Weimar Republic or Zimbabwe need very careful scrutiny. Hyperinflations are caused by quite specific circumstances, although there are some shared characteristics of countries and

monetary regimes that experience hyperinflation. While causes can be complex and varied, the Monetarist explanation sheds almost no light on the experience.

- Despite higher deficit ratios in the US, UK and Japan since the GFC, and higher stocks of bank reserves, which were caused by quantitative easing and the interest rate policies of the central banks, these countries have not endured hyperinflation or even high inflation. Nor is there any sign of this occuring in the foreseeable future.

Most critics of MMT and of so-called fiat money appear to imagine a past in which money was closely tied to a commodity such as gold, which constrained the ability of both government and banks to create money **out of thin air**. The best example was the precious metal coin that supposedly gave a real value to government money, and forced government to actually get gold in order to spend. A strict gold standard with 100 per cent gold backing against paper notes (issued by government or banks) accomplished the same task.

These critics would also advocate the formalisation of constraints on fiscal policy in the form of balanced budget amendments, debt limits, or for deficit doves, a commitment to eventually slash deficit spending once the recovery gets under way. These already exist in the USA. In 2015, the UK Parliament passed a bill requiring the UK Treasury to achieve fiscal surpluses after 2019–20 when GDP growth exceeded one per cent, subject to the absence of spending shocks. This was amended in late 2016 under the pretext of the negative impact of Brexit on private spending and hence fiscal outcomes.

As we have argued, a floating exchange rate provides policy space that can be used by prudent governments with their own sovereign currencies to pursue domestic policy goals with a greater degree of freedom. Except for the losers of the First World War (plus Poland and Russia, which were on the winning side but had left the capitalist world), there are no cases of nominally democratic Western capitalist countries that have experienced hyperinflation in the past century. And if we limit our data set to those with floating currencies, there were no countries with exchange rate crises either (see Chapter 24).

It is only countries with fixed exchange rates or other promises to deliver foreign currency or gold (such as debts in foreign currencies) that experience hyperinflations and currency crises. This tends to result from the imprudent expansion of these IOUs relative to the ability to actually deliver the foreign currency or gold. While it appears that the fixed exchange rate guarantees prudence, this is not achieved in practice. The fixed exchange rate introduces exchange rate crises plus involuntary default as possibilities, and does not guarantee that government will be prudent.

When a sovereign government promises to deliver foreign currency, it actually exposes the nation to Weimar Republic-style hyperinflationary risks. This risk can be compounded if their banks are not necessarily prudent.

Real world hyperinflations

In reality, high inflation and hyperinflation are rare events. In this section, we look at historical examples of hyperinflation periods. The simple explanation of **printing money** to finance excessive fiscal deficits sheds almost no light on these episodes.

Today, the best-known cases of hyperinflation occurred during the Weimar Republic and more recently in Zimbabwe. (Less well known, but more spectacular, was the Hungarian hyperinflation of 1946.) Mitchell (2010; 2011) provides an insightful analysis. We will not reproduce his work here, but will summarise key points about the Weimar and Zimbabwe hyperinflations to assure readers that these were not simple cases of too much 'money printing' to finance governments that were 'running amuck'.

The usual story told about Weimar Germany is that the government began to freely print a fiat money with no gold standing behind it, without regard for the hyperinflationary consequences. The reality is more complex. First, we must understand that even in the early 20th century, most governments spent by issuing IOUs, although many were convertible on demand to UK Sterling or gold. Having lost the First World War, Germany was suffering under the burden of impossibly large reparations payments that had to be made in gold, but it had very limited gold reserves. To make matters worse, much of its productive capacity had been destroyed or captured. Germany was supposed to export to earn the gold needed to make the payments demanded by the victors.

Keynes (1920) argued in his first globally famous book, *The Economic Consequences of the Peace*, that Germany could not possibly pay the external debts denominated essentially in gold.

The problems for the Weimar Republic thus began long before the hyperinflation emerged in 1923. The reparations payments required under the Versailles Treaty squeezed the German government so badly that it eventually defaulted. The French and Belgian armies then retaliated after the German default and took over the industrial area of the Ruhr, Germany's mining and manufacturing heartland. The Germans in turn stopped work, and production ground to a halt. The Germans also kept paying the workers in local currency despite limited production being possible.

The nation's productive capacity was not even sufficient to satisfy domestic demand, and so Germany was unable to export goods to earn gold to pay the huge reparations. The government believed that it was politically impossible to raise taxes to a sufficient level to match the value of resources needed for exports to pay the reparations. Instead it relied on spending in excess of taxes. This meant that government competed with domestic demand for a limited supply of output, thereby driving up prices.

At the same time, Germany's domestic producers had to borrow abroad (in foreign currency) to buy the imports that they needed. Rising prices plus foreign borrowing caused a depreciation of the domestic currency, which then increased the need to borrow (since foreign imports cost more in terms of domestic currency) and at the same time increased the cost of the reparations in terms of domestic currency.

The crunch came when the export trade stalled and the only way the German government could keep paying its treaty obligations was to keep expanding its fiscal deficit. The hyperinflation that followed was inevitable given these circumstances.

While it is often claimed that the Weimar central bank contributed to the inflation by purchasing debt from the treasury, actually it operated much like central banks do today: it sold government debt to banks, offering them a higher earning asset in exchange for reserves. Fiscal deficits grew rapidly from the high inflation, and then hyperinflation as tax revenue could not keep pace with rising prices.

Finally, in 1924 Germany adopted a new currency, and while it was not legal tender, it was designated as being acceptable for the payment of taxes. The hyperinflation ended. It is evident that the major supply constraints in the post-First World War period, plus the obligation to make substantial reparations, were key factors in the Weimar's hyperinflation.

Let us turn next to Zimbabwe, which is increasingly used to claim that continuous and rising government deficits will generate hyperinflation. A careful examination of Zimbabwe's experience does not allow us to make a case against the use of continuous fiscal deficits in defence of full employment.

White minority rule in Colonial Africa had created very unfair land sharing arrangements between the whites and majority blacks, and Zimbabwe (then known as Rhodesia) was no exception. Whites, who constituted one per cent of the population, owned 70 per cent or more of the productive land in Rhodesia.

After the civil war of the 1970s and gaining independence in 1980, Rhodesia became Zimbabwe and Robert Mugabe's government initially oversaw relatively improved growth with stable inflation outcomes. Apart from a severe drought in 1992–3 which temporarily increased the inflation rate, real GDP growth was relatively robust.

The problems came after 2000 when the Mugabe government introduced land reforms to speed up the process of reducing inequality. The revolutionary fighters who had achieved Zimbabwe's freedom from colonial rule were given the productive, white-owned commercial farms which had hitherto fed the population and were the largest employers. While these reforms were well motivated, the problem was that the new owners had no background in running commercial agriculture. The resulting collapse of food production was catastrophic. Potential output levels contracted sharply, with around 45 per cent of the food output capacity being destroyed. Unemployment rose above 80 per cent and underemployment became widespread. A rapid demand contraction was required, but the rising unemployment made this impossible to implement politically.

The independent government also mismanaged its public infrastructure and the constraints flowed through the supply chain. For example, the National Railways of Zimbabwe system was degraded and its capacity to transport mining export output fell substantially. In 2007, there was a 57 per cent decline in export mineral shipments.

The manufacturing sector was also affected. Manufacturing output fell by 29 per cent in 2005, 18 per cent in 2006 and 28 per cent in 2007. In 2007, only 18.9 per cent of Zimbabwe's industrial capacity was being used. This reflected a range of problems, including raw material shortages. But overall, the manufacturers blamed the central bank for stalling their access to the foreign exchange which was needed to buy imported raw materials. Instead, the Reserve Bank of Zimbabwe had diverted its foreign reserves into food imports to try to address the collapse in farm output.

The collapse in the supply side of the Zimbabwean economy explains the hyperinflation that followed. To avoid the hyperinflation, given the supply side contraction, the Zimbabwean government would have had to severely contract real spending to match the new lower capacity. Had the government implemented austerity to this degree, widespread starvation and deaths would have followed.

While the hyperinflation was almost inevitable in the circumstances, it provides no intrinsic case against a government that is sovereign in its own currency, that runs permanent deficits to pursue full employment, and where productive capacity grows over time.

A private sector investment boom, for example, would have caused the same outcome both in inflation and the political problems of fighting it. In general, any sustained aggregate spending expansion in the context of the aggregate supply contraction would have led to the same hyperinflationary outcome.

This was also another case in which government could not have raised taxes for both political and economic reasons. Again, to label this a simple Monetarist case of government **printing money** really sheds no light on Zimbabwe's problems, which were caused mostly by social unrest, the collapse of agriculture and hence major supply shortages, and consequently, heavy external debt.

Summing up on hyperinflation

It is important to acknowledge that greater constraints on government spending (or a greater capacity to increase taxes) might have successfully prevented hyperinflation in both cases. However, when these specific cases of hyperinflation are studied, it is evidently not a simple story of a government adopting a fiat money, and printing and spending too much. There are many paths to hyperinflation, but there are common problems: social and political upheaval, civil war, the collapse of productive capacity that could be due to war, weak government, and foreign debt denominated in an external currency or gold. In these circumstances, we do observe rising fiscal deficits and (by identity) growing outstanding government IOUs. But we also find banks creating money to finance private spending that competes with government to drive up prices.

Therefore, it is likely that tighter fiscal policy would have helped to reduce inflationary pressures. This probably would not have reduced overall suffering, since a common cause of hyperinflation is some kind of supply constraint on output. But the solution to the problems does not require the adoption of a gold standard. Rather, to tackle a problem of high inflation, policymakers should try to slow down the indexing of wages and transfers, stabilise production, reduce demand relative to supply, and quell social unrest. When high inflation has persisted for some time, it also helps to adopt a new currency and to default on external debts.

Thus, there is a link between high (or hyper) inflation, fiscal deficits, and money supply, although not a simple dynamic Monetarist relationship. A government always spends by **keystrokes** that credit accounts, and taxes (or sells bonds) by reverse keystrokes that debit accounts. As discussed above, in high or hyperinflation periods, taxes (debits to accounts) grow slower than government spending (credits to accounts) so we expect deficits to result, which means that outstanding government IOUs (the monetary base plus government debt) grow.

Matters are made worse if a high interest rate policy is pursued by the central bank. This is because government typically matches its rising deficit with new government bond (debt) issues, and interest payments add to government spending. If the central bank reacts to growing deficits by raising interest rate targets, it helps to fuel the growth of the deficit and also adds demand stimulus to the economy in the form of interest payments by government, but at the same time deters investment, which is counterproductive.

Conclusion

On one level, the MMT approach is descriptive: it explains how a sovereign currency works in practice. When we talk about government spending by keystrokes and argue that the issuer of a sovereign currency cannot run out of them, that statement is descriptive and factual.

Equally, the statement that sovereign governments do not borrow their own currency is also descriptive and factual. Our classification of bond sales as part of a monetary policy to help the central bank hit its interest rate target is also descriptive and factual. And finally, our argument that a floating exchange rate provides the most domestic policy space is also descriptive and factual.

Functional finance principles provide a framework for prescriptive policy. Thus sovereign government ought to operate fiscal and monetary policy to achieve full employment. In Lerner's view, this is done by setting the government's net spending at the right level, spending more and taxing less when there is unemployment, and setting the interest rate at the right level. That isn't very radical. The policy was adopted by post-Second World War Keynesians, and also even by Milton Friedman (who had his own version of functional finance).

However, Lerner's initial proposal was formulated in an economic environment of low inflation. Indeed, there were concerns about a return to deflation, such as that suffered in the 1930s. Later, after inflation reared its ugly head during the 1960s, Lerner became quite concerned about price stability. He developed a policy proposal which argued for wage and price controls. However, since the late 1970s the major countries have always relied on tight fiscal and monetary policy to fight inflation.

The problem is that governments then had to abandon any pretence that they were pursuing full employment. Indeed, unemployment became a tool for achieving price stability. Since the 1970s, the conventional wisdom has been that central banks ought to pursue only price stability, with the role of fiscal policy being downgraded. Lerner's 'steering wheel' approach to policy was abandoned. The result has typically been high unemployment and substandard economic growth. In the US, poverty and inequality have risen. Globally, growing unemployment has been a problem even during economic expansions.

In Chapter 19, we examined an alternative strategy to create jobs without sparking inflation – the Job Guarantee (JG) approach. The JG programme directly targets the unemployed to lift them out of poverty. The alternative approach of using general tax cuts or spending increases tends to favour the already relatively well-off rather than leading to new job opportunities **trickling down** to the unemployed and poor.

A sovereign currency needs an **anchor**, and by setting the basic wage in a JG programme, the programme itself becomes the anchor. As long as the programme wage is held steady, and so long as there are employees in the programme, an employer can recruit a new worker out of the programme at a slightly higher wage.

Operating the economy at full employment and with a relatively stable wage in a buffer stock jobs programme will help to stabilise not only consumption spending and household income, but it will also help to stabilise wages and therefore prices.

References

Cagan, P. (1956) "The Monetary Dynamics of Hyperinflation", in M. Friedman (ed.), *Studies in the Quantity Theory of Money*, Chicago: University of Chicago Press.

Friedman, M. (1953) *Essays in Positive Economics*, Chicago: University of Chicago Press.

Friedman, M. (1969) "The Optimum Quantity of Money", in *The Optimum Quantity of Money and Other Essays*, Chicago: Aldine Publishing Company.

Keynes, J.M. (1920) *The Economic Consequences of the Peace*, New York: Harcourt Brace.

Lavoie, M. (2013) "The Monetary and Fiscal Nexus of Neo-Chartalism: A Friendly Critique", *Journal of Economic Issues*, 47(1), 1–32.

Lerner, A. (1943) "Functional Finance and the Federal Debt", *Social Research*, 10(1), 38–51.

Lerner, A.P. (1947) "Money as a Creature of the State" *The American Economic Review, Papers and Proceedings*, 37(2), 312–21.

Mitchell, W.F. (2010) *Modern Monetary Theory and Inflation – Part 1*. Available at: http://bilbo.economicoutlook.net/blog/?p=10554, accessed 1 March 2017.

Mitchell, W.F. (2011) *Modern Monetary Theory and Inflation – Part 2*. Available at: http://bilbo.economicoutlook.net/blog/?p=13035, accessed 1 March 2017.

Mitchell, W.F. (2015) *Eurozone Dystopia: Groupthink and Denial on a Grand Scale*, Cheltenham: Edward Elgar.

Mitchell, W.F. and Juniper, J. (2007), "Towards a Spatial Keynesian macroeconomics", in P. Arestis and G. Zezza (eds), *Advances in Monetary Policy and Macroeconomics*, Basingstoke: Palgrave Macmillan, 192–211.

Endnotes

1. Strictly, a positive overnight interest rate can be targeted without debt sales by setting the target rate to the deposit rate paid by the central bank on reserves.
2. The IS-LM framework presented in Chapter 28 is typically used to substantiate this idea in mainstream analysis.
3. Repo is an abbreviation of 'repurchase agreement'. A holder of government securities sells securities to a lender and agrees to repurchase them at an agreed future date at an agreed price.
4. There is normally a positive term structure of yields since longer maturity bonds pay higher interest rates. An important reason for this is capital risk: if the central bank raises its target rate before maturity, the market price of longer-term bonds falls.
5. Government needs to pay the market wage across sectors and skill levels to employ workers either directly or indirectly to get projects that are funded by the 'pump priming' under way.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

22 FISCAL SPACE AND FISCAL SUSTAINABILITY

Chapter Outline

Learning Objectives

- Develop an understanding of the concept of fiscal space.
- Develop an understanding of the concept of fiscal sustainability.
- Be able to articulate why a floating exchange rate maximises fiscal space.
- Be able to articulate why a sovereign currency issuer that adopts a floating exchange rate can pursue a functional finance approach to budgeting.
- Recognise that a government operating with a sovereign currency will never face a crisis associated with public debt sustainability.

22.1 Introduction

In this chapter, we bring together the discussion in Chapters 20 and 21 to formulate a concept of **fiscal sustainability** that is consistent with the principles of MMT. We contrast this notion with the concept of **fiscal space**, which is a conceptual framework used by international agencies such as the IMF to assert that there are financial limits on the capacity of currency-issuing governments to run fiscal deficits.

We show that the only meaningful way in which the concepts of fiscal space and fiscal sustainability can be understood is in terms of real resource availability.

In designing good fiscal policy, it is crucial that governments understand what the limits on fiscal deficits are, given that terms such as fiscal sustainability, fiscal consolidation, fiscal austerity and budget repair are in the media almost every day.

The mainstream version of fiscal responsibility is based on the false premise that currency-issuing governments have limited financial resources and must maintain a policy stance that appeases the bond market lenders or risk 'running out of money'.

We show in this chapter that this reasoning is not applicable for sovereign governments. Rather, MMT provides a coherent **full employment fiscal deficit condition** that a responsible government would seek to meet – and whether that requires a larger deficit, a smaller deficit, or even under some conditions, a surplus, should never be a concern of government.

Fiscal sustainability within a MMT framework requires that the government uses its fiscal capacity to sustain full employment and price stability rather than being diverted by targets specified purely in terms of some notional public debt-to-GDP ratio, or a deficit-to-GDP ratio.

Finally we use a technical exposition to refute the claim that reliance on fiscal policy to sustain full employment is likely to lead to a rising treasury debt-to-GDP ratio, which eventually becomes unsustainable.

22.2 The Full Employment Fiscal Deficit Condition

We begin with the observation that in an open economy, if there were no government spending or taxation, the level of economic activity (output) would be determined by private domestic spending (consumption plus investment) and net external spending (exports minus imports). If one or more of those components of spending declines then activity will decline.

A **spending gap** occurs if total spending is less than that required to support output levels, which at current productivity levels will provide enough jobs (measured in working hours) for all the workers who desire to work. In other words, at full employment there is no spending gap. From a functional finance perspective, the role of government fiscal policy therefore is to ensure there is no spending gap.

It becomes obvious (and incontestable) that if non-government spending declines from a level associated with full employment, a spending gap emerges and the only way it can be eliminated is via intervention to increase net spending (by direct government spending and/or a tax cut to encourage non-government spending).

In terms of the expenditure output approach outlined in Chapter 15, we know that the sources of spending flows which add to aggregate demand are:

Household consumption (C)

Private investment (I)

Government spending (G)

Export revenue (X)

(Note that we are abstracting from net income flows and transfers on the current account in this section.)

The income (payments to resource owners involved in the production of output) that is generated by these spending flows can be used in the following ways:

Taxation payments net of transfers (T)

Household consumption (C)

Household saving (S)

Import spending (M)

Clearly, the sources of income must be equal to its uses, which is a convention that underpins the National Accounts.

This allows us to write the two sides of income generation like this:

$$C + I + G + X = C + S + T + M \tag{22.1}$$

Given C cancels out we know that:

$$I + G + X = S + T + M \tag{22.2}$$

The left-hand side of Equation (22.2) represents the exogenous **injections** into the spending flow, and on the right-hand side are the **leakages** from the income stream. Injections constitute new spending in the economy, whereas the leakages are induced by changes in national income and drain aggregate spending.

The two sides of Equation (22.2) are brought into equality by national income adjustments (that is, variations in the level of aggregate activity brought about by spending variations).

For example, if I increases (with G and X constant), the increase in aggregate spending stimulates firms to increase output to meet the new orders. In doing so, they increase employment and the increased income sets off the expenditure multiplier process which we explained in Chapter 15.

Income recipients use the extra income to increase saving (S), pay more tax (T) (even if tax rates are unaltered), and buy more imports (M), in addition to higher consumption.

The economy stops expanding again once the change in investment (the injection, holding G and X constant) is equal to the sum of the changes in S, T and M (leakages). The new equilibrium position (the point when income stops expanding) thus occurs when the sum of the injections equals the sum of the leakages. Whenever a position of rest (equilibrium) is disturbed (by a spending injection), national income adjusts and brings the total income sensitive leakages, the right-hand side of Equation (22.2), into line with the new total level of injections, the left-hand side of Equation (22.2). At that point the system is at rest.

Three points should be noted. First, this position of 'rest' does not necessarily have to coincide with full employment. The system will adjust to dramatically lower levels of injections and come to rest even if there are high unemployment levels. Keynes (and Marx before him) demonstrated that economies could come to rest at very high levels of unemployment and would remain in that state unless budged by an intervention (that is, a fiscal policy stimulus).

Second, when an economy is at 'rest' and there is high unemployment, there must be a spending gap, given that mass unemployment is the result of deficient effective demand. With weak non-government spending, the injection via increasing government spending (G) must more than offset the increased drain (leakage) from taxation revenue, net of transfers (T). That is, a fiscal deficit is needed because there is a non-government spending gap.

Third, this doesn't mean that a fiscal deficit is always required. We need one more condition to establish the case for ongoing fiscal deficits. If the non-government decisions taken together (so consumption and saving decisions by households, investment decisions by production firms, and the balance of the external sector) indicate a desire to 'net save', that is, $I + X < S + M$ at a particular level of economic activity, then the only way that level of activity can be maintained on an ongoing basis (at any rate of unemployment) is if $G > T$. That is, a fiscal deficit is required on a continuous basis to sustain a given level of activity.

In this case, a fiscal deficit 'finances' the desire by the non-government (domestic and foreign) sector to save overall by maintaining sufficient demand to produce a level of income which will generate the desired level of net saving.

Functional finance tells us that responsible fiscal policy requires two conditions to be fulfilled. First, the fiscal position (deficit or surplus) must fill the gap between the savings and investment minus the gap between exports and imports.

Thus:

$$(22.3) \quad (G - T) = (S - I) - (X - M)$$

For national income to be in equilibrium, the fiscal deficit $(G - T)$ will equal the excess of saving over investment (which drains domestic demand) minus the excess of exports over imports (which adds to demand).

If the right-hand side of the equation, $(S - I) - (X - M) > 0$, is in surplus overall, we know that the non-government sector is net saving overall, and the only way the level of national income can come to rest at this level is if the fiscal deficit offsets that surplus. Note that this concept of net saving is broader than household saving out of disposable income.

A surplus on the right-hand side of Equation (22.3) can arise from $(S - I) > (X - M)$ (that is, the private domestic sector net saving being more than the net export surplus) or it could be associated with a net exports deficit (draining demand and adding foreign savings) being greater than the private domestic sector deficit (investment greater than saving) which adds to demand.

Second, we have noted that an equilibrium level of national income may not simultaneously satisfy our desire for full employment. We can define a full employment level of national income as that which is generated when all resources are fully utilised according to the preferences of workers and owners of other productive resources.

Given that S, T and M are all positively related to the level of national income, there is a unique level of each of these flows that is defined at full employment. Changes in behaviour (for example, an increased desire to save per dollar earned) will change that 'unique' level but for given behavioural preferences and parameters, we can define full employment levels for each.

We define $S(Yf)$ and $M(Yf)$ as being the saving and import flows that would occur when national income is at its full employment level (Yf). We also consider investment to be sensitive to national income (we discuss the accelerator theory of investment in Chapter 25) such that a higher level of output gives rise to an increased inducement to invest for a given technology. So $I(Yf)$ might be defined as the full employment flow of investment. We consider export spending to be determined by the level of world income.

Finally, we note that taxes net of transfers (T) are sensitive to the economic cycle (the automatic stabiliser component of fiscal policy). Thus, we can define $T(Yf)$ as being the level of net taxes that the government receives when the economy is at full employment, given current tax rates and transfer policy settings.

Accordingly, the condition for equilibrium national income that will also sustain full employment is written as:

$$(22.4) \qquad [G - T(Yf)] = S(Yf) + M(Yf) - I(Yf) - X$$

Equation (22.4) is what we call the **full employment fiscal condition**.

The sum of the terms $S(Yf)$ and $M(Yf)$ represents drains on aggregate demand when the economy is at full employment, and the sum of the terms $I(Yf)$ and X represents spending injections at full employment.

If the drains on the right-hand side of Equation (22.4) outweigh the injections on the right-hand side of Equation (22.4), then full employment will be achieved if the fiscal deficit $[G - T(Yf)]$ is sufficient to offset the non-government spending gap. If the fiscal deficit is not sufficient, then full employment will not be achieved.

On the other hand, if aggregate spending exceeds the value of full employment output, then inventories will be depleted and prices will tend to increase. If government expects that total demand will continue to exceed productive capacity (and thus, generate continuing increases in the price level), then it may decide to cut net government spending to reduce inflation pressures.

In this sense, the principles of MMT specify a strict discipline on fiscal policy. If the goal is full employment and price stability, then the full employment fiscal deficit condition must be met.

22.3 Fiscal Space and Fiscal Sustainability

The discussion in the previous section allows us to understand what fiscal sustainability means. The concepts of fiscal sustainability and fiscal space are well developed in the economics literature. The problem is that in general they are based on the flawed understanding that sovereign governments operate with financial constraints that are beyond their control.

The IMF (2005) defines fiscal space as:

> *[the] room in a government's budget that allows it to provide resources for a desired purpose without jeopardizing the sustainability of its financial position or the stability of the economy. The idea is that fiscal space must exist or be created if extra resources are to be made available for worthwhile government spending. A government can create fiscal space by raising taxes, securing outside grants, cutting lower priority expenditure, borrowing resources (from citizens or foreign lenders), or borrowing from the banking system (and thereby expanding the money supply). But it must do this without compromising macroeconomic stability and fiscal sustainability – making sure that it has the capacity in the short term and the longer term to finance its desired expenditure programs as well as to service its debt.*

This IMF definition or variations of it are used widely by government agencies and other multilateral institutions to place limits on fiscal policy.

The definition is both narrow and vague in that it considers fiscal sustainability to be indicated by inexact thresholds for financial indicators such as the fiscal deficit-to-GDP ratio or the public debt-to-GDP ratio, rather than placing the fiscal outcome in the wider context of the overall state of the economy. There has never been agreement or coherent research to define some public debt ratio past which governments cease to be viable.

Most importantly, the definition assumes that the currency-issuing government has the same financial constraints as a firm or household, and ultimately, its ability to spend and service its outstanding financial liabilities is dependent on whether bond markets will continue to loan it funds at realistic interest rates.

As we have learned, in a fiat monetary system, these concepts of fiscal space ignore two key points:

- A sovereign government is not revenue constrained, which means that fiscal space cannot be defined in financial terms.
- The capacity of the sovereign government to mobilise resources depends only on the **real** resources (labour, natural resources, productive capacity, and know-how) available to the nation.

Given that a sovereign government is never in danger of running out of money and is thus always solvent (unless it chooses for political reasons not to meet promised payments), the concept of fiscal sustainability (and fiscal space) will not be defined by the standard financial ratios (for example, public debt ratios) on which organisations such as the IMF choose to focus.

From an MMT perspective, the sovereign government's overriding responsibility is to pursue public purpose and well-being through fiscal sustainability, which must be constructed in terms of maintaining full employment and price stability. The *full employment fiscal deficit condition* provides the focus for policymakers in this regard.

Several considerations then become important in our concept of fiscal sustainability.

Advancement of public purpose

Once the non-government sector has made its spending (and saving) decisions based on its expectations of the future, the government must render those private decisions consistent with the objective of full employment.

Given that the non-government sector will typically desire to net save (accumulate financial assets in the currency of issue) over the course of an economic cycle, this means that there will be on average a spending gap over the course of the same cycle that can only be filled by the national government. A non-government desire to run a surplus will be achieved only if the government runs a deficit.

The national government then should seek to satisfy the *full employment fiscal deficit condition*. We might call these deficits 'good' because they are associated with levels of economic activity that support non-government saving preferences and maintain jobs for all who desire to work.

By way of contrast, it is possible that the government's political goals may be to maintain some slack in the economy (persistent unemployment and underemployment) which means that the government deficit will be somewhat smaller and perhaps even, for a time, a fiscal surplus will be possible. But adopting this strategy would introduce fiscal drag (deflationary forces) into the economy which will ultimately cause firms to reduce production and income and drive the fiscal outcome towards increasing deficits via lost tax revenue and increased welfare payments. In other words, trying to reduce a government deficit can actually slow the economy by so much that a larger deficit is actually generated.

Ultimately, the spending gap will be closed by the automatic stabilisers because falling national income ensures that the leakages (saving, taxation and imports) equal the injections (investment, government spending and exports) so that the sectoral balance identity, which is an accounting construct, will hold. But at that point, the economy will support lower employment levels and higher unemployment. The resulting fiscal deficit is what we would call a 'bad' deficit because it has been driven by a declining economy and rising unemployment.

Fiscal sustainability thus requires that the government runs 'good' deficits consistent with full employment. Any fiscal strategy that runs counter to that goal is unsustainable in the sense that it is not consistent with full employment.

Note that we are presuming that the 'normal' case is for the non-government sector taken as a whole to desire a surplus so that the government normally runs a deficit. However, a few nations run sufficiently large current account surpluses that their 'normal' government balance and their domestic private sector balance are both in surplus.

Understanding the monetary environment

Any notion of fiscal sustainability must be related to the intrinsic nature of the monetary system within which the government is operating. It makes no sense to comment on the behaviour of a government in a fiat monetary system using the logic that applied to a government operating under a gold standard, where the currency was convertible to another commodity of intrinsic value and exchange rates were fixed.

The constraints that the latter monetary system imposed on the national government – the necessity of constraining fiscal and monetary policy to defend the exchange parity – **do not** apply to a national government in a fiat monetary system.

A government operating in a fiat monetary system may adopt voluntary restraints that allow it to replicate the operations of a government during a gold standard. These constraints may include matching deficits with public debt issuance; setting ceilings on the size of deficits; limiting the growth in government spending over some finite period; and placing a ceiling on how much public debt can be outstanding.

None of these financial constraints have any applicability to a sovereign government operating in a fiat monetary system. In general, the imposition of these restraints reflects ideological imperatives which typically involve a desire to reduce the importance of government in the economy.

Accordingly, the concept of fiscal sustainability in the case of a sovereign government does not include any recognition of the legitimacy of these voluntary restraints. These constraints essentially mean that the responsibility of the national government to pursue full employment is subjugated to meeting meaningless financial constraints.

Note that the decision to give up a sovereign, floating exchange rate, currency in favour of a gold standard or a fixed exchange rate system, is itself, a voluntarily imposed restraint.

Understanding what a sovereign government is

As we have learned, a sovereign government can spend whenever it wants and has no imperative to seek funds from the non-government sector to facilitate that spending. This is in sharp contradistinction with non-government entities which use the currency of issue. These entities can only spend after they have earned income, run down saving, borrowed, and/or sold assets. Such an entity cannot run permanent deficits. These constraints can never apply to a sovereign government in a fiat monetary system for reasons we have already outlined.

Defining fiscal sustainability in terms of the desirability of fiscal surpluses overlooks the reality that if the government runs a surplus, then the non-government sector must be in deficit. Or, to put it the other way, if the non-government sector is to run a surplus, the government must be in deficit.

In nations without external surpluses, a government surplus will always be reflected in a private domestic deficit. This cannot be a viable growth strategy because the private sector (which faces a financing constraint) cannot be in deficit (and accumulating ever-increasing debt) on an ongoing basis. Ultimately, the fiscal drag will force the economy into recession as the private domestic sector seeks to save again to reduce the precariousness of its debt exposure.

Accordingly, the concept of fiscal sustainability involves a conceptualisation of a government which is free of financial constraints and has a range of possibilities that are not available to any non-government entity.

Further, given that the non-government sector will typically net save in the currency of issue, a sovereign **government** must **run** deficits on a continuous basis. The proper size of those deficits will be dictated by the *full employment fiscal deficit condition*.

Understanding why governments tax

We learned in Chapters 20 and 21 that taxation functions to promote offers from the non-government sector of goods and services to the government sector in return for the necessary funds to extinguish legally enforceable tax liabilities.

In this way, the imposition of taxes creates unemployment (people seeking paid work) in the non-government sector. We might think of this as taxes creating the real resource space (by reducing non-government purchasing power) that the government has to fill with government sector spending. The real resource space created by the taxes allows a transfer of goods and services from the non-government to the government sector, which in turn, facilitates the government's economic and social programme.

We can think of the effect of imposing taxes as creating unemployed resources (including labour). Government spending then puts those resources to use in the public sphere. It would make no sense for government to impose taxes, causing unemployment, except to the extent that it needs to release resources from private use so that they can be employed in the public sector.

The funds necessary to pay the tax liabilities are provided to the non-government sector by government spending. Accordingly, government spending provides the paid work which eliminates the unemployment created by the taxes. If $G < T$, then the non-government sector is squeezed for liquidity and must retrench wealth. A recession typically follows.

Unemployment occurs when the non-government sector desires to earn the monetary unit of account, but doesn't desire to spend all it earns, other things being equal. As a result, involuntary inventory accumulation among sellers of goods and services translates into decreased output and employment.

The obvious conclusion is that unemployment occurs when net government spending is too low to accommodate the non-government sector's need to pay taxes and its desire to save some of its overall income.

Accordingly, the concept of fiscal sustainability does not entertain notions that the continuous deficits required to finance non-government net saving desires in the currency of issue will have to be 'paid back' via the imposition of higher taxes in the future.

A fiscal deficit in one period is not paid back in some later period. Taxes in the future might be higher or lower or unchanged depending on the overall state of aggregate spending in relation to productive capacity. These future decisions about tax rates have nothing to do with 'funding' government spending.

Understanding why governments issue debt

We have also learned that a sovereign government does not have to issue debt to the non-government sector before it can spend. In Chapters 20 and 21, we explained the way in which public debt is used as a vehicle to allow the central bank to stabilise the interest rate around its desired target rate (through open market operations).

In fact, a sovereign government does not need to issue debt at all. The central bank could still maintain a non-zero policy interest rate by simply paying a competitive rate to the commercial banks on any excess reserves that ongoing fiscal deficits had created in the banking system.

In other words, the practice of matching of fiscal deficits with debt issuance to the non-government sector is not essential, and reflects past habits and modern ideological views about limits to be placed on government spending. It is thought that if governments are forced by institutional arrangements to issue debt to accompany every spending choice, then the public scrutiny of the public debt build-up will put a brake on the spending. That is why the public debate is dominated by a focus on public debt ratios, when in fact they are of no importance in determining whether a government is fulfilling its responsibilities to advance public purpose.

Accordingly, the concept of fiscal sustainability should never make any financing link between debt issuance and net government spending. There is no inevitability for debt to rise as deficits rise. Voluntary decisions by the government to make such a link have no basis in the fundamentals of the fiat monetary system.

Setting fiscal targets

Setting fiscal targets is a fraught exercise for governments because we know that the final fiscal outcome is dependent on both the discretionary spending and tax parameters that the government sets and the spending choices of the non-governmental sector.

Rather than define fiscal sustainability in terms of some hard-and-fast public debt thresholds, deficit limits or even spending limits, the government should rather focus on maintaining net public spending at levels which sustain aggregate demand at its full employment level. If it achieves that real target (full employment), then the fiscal balance will move accordingly and be sustainable.

Foreign exposure

It is often thought that a nation is exposed to credit risks if foreigners purchase the debt issued by the national government. This fear is, however, misplaced.

In Chapter 24, we will see that a nation must be prepared to export more than it imports if it desires to accumulate financial assets issued by the deficit nations. For example, China runs external surpluses against the US in goods and services because it desires to build up a wealth portfolio of financial assets denominated in US dollars. That is the only way a foreign nation can accumulate net wealth in another nation's currency.

The foreign wealth holders might choose to hold some of that wealth in the form of US government bonds, say. But that doesn't mean that US government spending is dependent on the preferences of these foreign bond holders. A sovereign government can always spend whether foreign bondholders choose to hold their trade surpluses in the form of that government's debt or not. Fiscal sustainability – the capacity to spend and transform real goods and services – is not compromised by foreign purchases of a nation's public debt, nor by a refusal to purchase the deficit nation's public debt.

Understanding what a cost is

The IMF concept of fiscal space reflects a misunderstanding as to what constitutes an economic cost. When a government announces in its fiscal statement that it will spend $100 billion on a new project, that number does not reflect a cost.

The real cost of any programme is the extra real resources that the programme requires for implementation. For example, if the government announced the introduction of a Job Guarantee, the real costs of the programme would be the extra consumption that the formerly unemployed workers could enjoy and the extra capital equipment that the workers would use in their productive pursuits. In other words, government programmes have to be appraised by how they use real resources rather than in terms of the nominal money values involved.

Accordingly, the concept of fiscal sustainability should be related to the utilisation rates of real resources, which takes us back to the initial point about the pursuit of public purpose: the correct fiscal balance is the one that is consistent with full employment.

22.4 The Debt Sustainability Debate

The discussion of government deficits and debts usually turns to the **sustainability** of continuous deficit spending that adds to debt, and possibly to the debt-to-GDP ratio as well. In this section, we will examine these issues. However, we will also argue that the modelling exercise is fundamentally misguided for a sovereign currency-issuing government, on two grounds.

First, **affordability *per se*** cannot be an issue for a sovereign government, and neither can sustainability in the sense that government can always make payments as they come due, no matter how large they become (see Chapter 21). However, if the debt to GDP ratio continuously grew, and interest payments on the debt grew faster than national income, while affordability cannot be an issue, the **crowding out** of other types of important government spending would be a concern.

Second, and equally important, is that the simple modelling framework of the growth process is flawed because it ignores the likelihood of changes to economic behaviour that would alter the dynamic relationship between the deficit and debt ratios.

Let us begin with a typical model used to evaluate the sustainability of deficit spending. At the outset, we should again note that orthodox economists view the **Government Fiscal Constraint** as an *ex ante* planning instrument and believe that the financing of deficits by the issue of the monetary base is inflationary. These arguments have been thoroughly canvassed and rejected by MMT in Section 21.3. However, to maintain consistency with the literature, all deficits lead to debt issue in the following model.

We measure the change in the debt-to-GDP ratio between two periods.

The debt-to-GDP ratio in period zero is given by $d_0 = D_0/Y_0$, where D_0 and Y_0 are debt and GDP, respectively.

Ceteris paribus, the outstanding debt in period zero (D_0) will grow at a rate r, due to interest payments which we assume for the moment to be constant (that is, to $D_0(1 + r)$). However, it will be offset(increased) to some extent by a government primary surplus(deficit) in period one (FS_1), which is defined as government spending less tax revenue, ignoring tax payments. The debt in period one is therefore given by $D_1 = D_0(1 + r) - FS_1$

Similarly, the level of GDP in period zero (Y_0) changes over period one (Y_1). We let g stand for growth of GDP, and we assume for the moment that g is constant. GDP grows to $Y_0(1+g)$ over period one, which could be sufficient to offset any rise in the debt level, so that the debt-to-GDP ratio remains constant or falls. These key variables, namely debt, the level of GDP and the government primary surplus, are usually in inflation adjusted terms, but that really does not matter; we can keep it all in nominal terms since 'deflation' by the inflation rate merely reduces all terms by the inflation rate. The change in the debt-to-GDP ratio between period zero and period one, Δd, is then $D_1/Y_1 - D_0/Y_0$. Substitution and careful algebraic manipulation yield the final expression for the change in the debt ratio between period zero and period one:

$$\text{(22.5)} \quad \Delta d = D_1/Y_1 - D_0/Y_0 = (D_0(1 + r) - FS_1)/(Y_0(1 + g)) - D_0/Y_0$$

$$= (D_0(r - g))/(Y_0(1 + g)) - FS_1/(Y_0(1 + g))$$

$$= (D_0/Y_0)(r - g)/(1 + g) - FS_1/Y_1$$

$$= d_0(r - g)/(1 + g) - fs_1$$

where fs denotes the surplus-to-GDP ratio.

Thus, according to this analysis, two factors assume importance in determining whether the debt ratio increases from one period to the next, namely:

- Whether the difference between the real interest rate on debt (r) and the real growth of GDP (g) is positive or negative; and
- Whether or not the government is running a primary surplus in period one, that is, $FS_1 > 0$.

Then there are four possible combinations of these two factors:

1. $r > g$ and the government runs a primary deficit
2. $r > g$ and the government runs a primary surplus
3. $r < g$ and the government runs a primary deficit
4. $r < g$ and the government runs a primary surplus

We can say that if $r > g$, then the debt ratio would continue to rise if the annual primary fiscal balances are zero or in deficit (Case 1).

In Case 2, if the government runs a constant primary surplus-to-GDP ratio, say fs^*, then there is a critical value of the debt ratio, say d^*, at which the debt ratio will remain constant, as long as r, g and fs^* remain constant. d^* can be written as:

$$\text{(22.6)} \quad d^* = fs^*(1 + g)/(r - g)$$

But this equilibrium debt ratio, d^*, is unstable. If for example, the primary surplus ratio temporarily departed from its previous value (fs^*), the debt ratio departs from its magnitude, d^*, and does not revert to d^*, even if fs is restored to its previous value, fs^*.

If the treasury runs a persistent deficit, but the real growth rate exceeds the real interest rate, so that $g > r$ (Case 3), then there is a constant debt ratio which satisfies Equation (22.6), with both the numerator and denominator now negative, so $d^* > 0$. Further it can be shown that, in contrast to the previous scenario, this equilibrium debt ratio is stable. Conversely, if there is both a primary surplus and $g>r$, (Case 4) then the debt ratio steadily declines.

Orthodox economists draw on a model of the type shown in Equation (22.5). Typically, they solve the model based on constant values of the key parameters, namely g, r, and fs, as we have done above. However, they fail to acknowledge that the growth rate of GDP (g) and the treasury's financial balance (fs) are interdependent in the presence of the **spending multiplier effect**.

Further, while the central bank has a major role in setting the overnight rate, which provides the base rate for the yield curve, it also has the capacity to flatten the yield curve by being prepared to buy longer-term debt and hence influence the interest rate on debt, r. Through quantitative easing, the USA, UK, Japan, and the Eurozone countries through the ECB have all been trying to stimulate private sector spending through lower long-term rates.

Thus, the levels of g, r and fs are not predetermined. This will have a major impact on the dynamics of the relationship between deficit and debt ratios.

We now examine several scenarios, which challenge the commonly held view that fiscal policy geared to the achievement of sustained full employment is likely to lead to an unsustainable debt ratio.

- Modest inflation tends to increase tax revenues through bracket creep, so that they grow faster than government spending, thus lowering deficits. Many would point to the tendency to generate 'negative' real interest rates, which makes it easier for real growth to exceed the real negative interest rate. In other words, the growth rate will again rise above the interest rate, and reverse the dynamics.
- Government can try to apply a stimulus to the macroeconomy by adjusting its fiscal stance (decreasing taxes and raising spending). Under plausible values of the expenditure multiplier and the initial debt ratio, the debt ratio will decline, due to GDP growing faster than total debt. By setting a low policy rate, the central bank will increase the prospect of a decline in the debt ratio, which will change the debt dynamics.
- The private sector may adjust its flows (spending and saving) in response to the government's fiscal stance. If government continually spends more than its income, it will be adding net wealth to the private sector; and its interest payments will add to private sector income. This increase in net wealth is likely to lead to additional private sector spending, and a lower saving ratio due to the **wealth effect**. Thus, private sector spending will rise relative to its income. The likely result is that tax revenues and consumption will rise, the private sector's surplus will fall and the government's deficit will fall. The explosive debt growth scenario is implausible and is based on the presumption that the non-government sector never changes its behaviour in reaction to larger government deficits that create more net financial wealth for the private sector.
- Finally, and this is the most contentious point. Suppose none of the dynamics just discussed come into play, so the government's debt ratio rises on trend. Will a sovereign government be forced to miss an interest payment? In the USA, the former Federal Reserve Chairman Bernanke explained that all central bank spending and lending to bail out the failing Wall Street investment banks during the GFC occurred by using keystrokes, or electronic entries on balance sheets. There is no technical or operational limit to its ability to do that.

We can conclude that there is a fundamental difference between perpetual private sector deficit spending and perpetual sovereign government sector deficits: the first really is unsustainable while the second is not.

We have argued that persistent government fiscal deficits lead to increasing private wealth and possibly higher treasury debt ratios (Watts and Sharpe, 2013). However, the sovereign currency-issuing government can continue to make all payments as they come due, no matter how big the debt ratio becomes. The act of making those payments *could* lead to inflation. It *could* lead also to policy changes, such as a lower interest rate target. It would probably also generate behavioural changes by the non-government sector which are likely to cause changes in growth rates, and deficit and debt ratios. Hence, a rising treasury debt ratio is unlikely to last forever.

As Lerner argued (see Chapter 21), we need to take a functional approach to fiscal policy. Rather than worrying about deficit and debt ratios, we should focus on what really matters: employment, growth, inflation, exchange rates, environmental sustainability, inequality, and other social and economic indicators of the quality of life.

Conclusion

In this chapter, we have defined a strict condition that a government should pursue in terms of its fiscal policy, the **full employment fiscal deficit condition**. Satisfying this condition ensures that aggregate spending is sufficient to sustain full employment.

In this sense, the sovereign government uses its fiscal capacity to support (finance) the saving desires of the non-government sector and ensure that there are no spending gaps in the economy, which ultimately would generate mass unemployment and recession.

Fiscal sustainability within a MMT framework requires that the government uses its fiscal capacity to sustain full employment and price stability, rather than be diverted by targets specified purely in terms of some notional public debt-to-GDP ratio, or a deficit-to-GDP ratio.

Finally a sovereign government is not constrained in its pursuit of sustained full employment by adverse deficit debt dynamics because it is never subject to a financing constraint.

References

IMF (2005) "Back to Basics – Fiscal Space: What It Is and How to Get It", *Finance and Development*, 42(2).

Watts, M.J. and Sharpe, T. (2013) "The Immutable Laws of Debt Dynamics", *Journal of Post Keynesian Economics*, 36(1), 59–84.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

23 MONETARY POLICY IN SOVEREIGN NATIONS

Chapter Outline

Learning Objectives

- Review arguments for interest rate rather than monetary targets by central banks.
- Develop the knowledge required to analyse the nature of liquidity management under different interest rate setting arrangements.
- Understand how changes in monetary policy are transmitted and impact on the macroeconomy.
- Recognise that the independence of the central bank is somewhat circumscribed.

23.1 Introduction

In Chapter 20 we examined monetary and fiscal operations. The central bank handles the treasury's tax revenue and receipts from bond sales as well as the treasury's payments. These fiscal operations necessarily impact on banking system reserves. Thus the central bank must work closely with the treasury in an accommodative manner to minimise fluctuations in bank reserves to ensure that the interest rate target is met. Through its examination of this coordination, MMT has made an important contribution to our understanding of these operational realities.

In Chapter 21, we explored the debates about the merits of a treasury engaging in discretionary fiscal policy. The claims that when 'financed' by monetary expansion, expansionary fiscal policy is inflationary and that financing by debt issue causes crowding out of private sector expenditure are rejected by MMT. Also the argument that like the private sector, a treasury could be subject to adverse deficit debt dynamics was rejected for a range of reasons in Chapter 22, including the reason that a sovereign issuer of the currency can always pay off debt denominated in its own currency.

In this fourth chapter on macroeconomic policy, our first objective is to bring together the earlier discussions of modern banking operations in the form of a short summary. We then analyse the switch, made by many central banks, from money supply targeting to interest rate targeting. We will then consolidate our understanding of liquidity management, that is, the central bank operations which ensure that the target interbank rate is achieved, by the consideration of different institutional settings. We will investigate the conduct of monetary policy by sovereign governments, which has been the dominant focus of macroeconomic policy in developed economies since the late 1970s. We will then examine and assess the adoption of unconventional forms of monetary policy by some central banks when they had limited capacity to ease (conventional) monetary policy any further.

We also briefly examine the unemployment and inflation outcomes for Australia, with reference to when inflation targeting was introduced in mid-1993, and then assess the advantages and disadvantages of adopting monetary policy as the primary macroeconomic policy tool.

The nature of central bank independence is investigated and this is followed by an integration of the verticalist and horizontalist strands of argument.

23.2 Modern Banking Operations

Private IOUs (funds owed) are denominated in the domestic currency. Similarly, private issuers of IOUs also promise to accept their own liabilities. For example, if a household has a loan with its bank, it can always pay principal and interest on the loan by writing a cheque on its deposit account at the bank. In this case, the bank accepts its own IOU in payment, just as governments accept their own liabilities (currency) in payments on debts due to government (tax liabilities).

Indeed, modern banking systems operate a cheque clearing facility so that each bank accepts cheques drawn on all other banks in the country. This allows anyone with a debt due to any bank in the country to present a cheque drawn on any other bank in the country for payment of the debt. The cheque clearing facility then operates to settle accounts among the banks. Banks clear accounts using government IOUs, and for that reason either keep some currency on hand in their vaults, or more importantly maintain **reserve deposits** at the central bank.

As we noted in Chapter 10, all modern financial systems have developed payments systems (of which the cheque clearing facility is a part) that ensure banks can get additional currency and reserves as necessary from other banks (called the interbank overnight market), or through borrowing them from the central bank. This access to funds enables them to clear accounts among themselves and with their depositors.

Also, banks try to minimise their holdings of currency in their vaults, not only due to security considerations, but more importantly because they would prefer to hold assets in the form of loans on which they will receive interest from the borrowers. Thus they leverage their currency reserves, and hold a small fraction of their assets in the form of reserves. So long as only a small percentage of their account holders try to convert deposits to cash on any given day, this is not a problem. However, in the case of a bank run (in which a large number of depositors try to convert on the same day), the bank will have to obtain currency from the central bank.

This can even lead to a **lender of last resort** action by the central bank when it lends currency reserves to a bank facing a run. In such an intervention, the central bank lends its own IOUs to the banks in exchange for their IOU, the bank gets a reserve credit from the central bank (an asset for the bank) and the central bank holds the bank's IOU as an asset. When cash is withdrawn from the bank, its reserves at the central bank are debited, and the bank debits the account of the depositor who withdrew funds. The cash then held by the depositor is the central bank's liability, offset by the bank's liability to the central bank.

23.3 Interest Rate Targets versus Monetary Targets

For a long time after the 1960s, Monetarist thinking, which focused on a central bank's purported ability to control the money supply, held sway in policymaking circles. Mainstream (orthodox) economists used to argue that the central bank could use quantitative constraints as a means of controlling the private creation of money (credit creation). Many central banks, including those in the USA, UK, Canada, Germany and Australia, targeted a monetary aggregate that is a measure of the **money supply** in the late 1970s and early 1980s. They did so because through the Quantity Theory of Money (QTM), the growth of the money supply in the long run was alleged to determine the inflation rate. However, by the mid-1980s they had all discovered that they were unable to control the money supply, and abandoned this major plank of Monetarist thinking.

Central banks now pay little attention to the growth rates of monetary aggregates, realising that they can really only set the price (typically the overnight interest rate), which has only an indirect impact on the quantity of reserves and the quantity of privately created money – and hence the money supply.

REMINDER BOX

Students may wish to review the discussion of the contrasting perspectives about the credit creation process and the **endogeneity of the money supply** in Chapter 10.

Despite central banks ceasing to target a monetary aggregate, monetary policy in the form of interest rate targeting has remained the primary arm of macroeconomic policy, and is still designed to control inflation. However, it is rarely explained how discretionary changes in monetary policy through changes in the target interbank rate can effectively target the inflation rate. At best, interest rate changes may lead to changes in spending, which could impact on the inflationary process.

The primacy of monetary policy over fiscal policy reflects the propositions arising from the expectations augmented Phillips curve, which became the dominant theory of inflation in the early 1970s. Fiscal policy was rejected as a strategy to achieve a lower rate of unemployment because of the inflationary consequences. Low and stable inflation, as opposed to full employment, became the main target of macroeconomic policy, reflecting the belief that an economic environment of low and stable inflation was the most conducive to private sector spending and employment because private sector planning would not be undermined by unexpected changes in the inflation rate. In addition, there was the presumption that an independent body charged with the formulation of monetary policy would make better decisions.

Some central banks are subject to formal **inflation targets**. For example, in 2018 the Bank of England was subject to a target inflation rate of two per cent per year. On the other hand, the Reserve Bank of Australia (RBA) adopted a **target range** and was required to maintain the Australian inflation rate at between two and three per cent per annum. The US Federal Reserve is not subject to a formal inflation target, but in a 2016 statement about longer-run policy goals and its monetary policy strategy, the Federal Open Market Committee (FOMC) stated that a CPI inflation rate target of two per cent was "most consistent over the longer run with the Federal Reserve's statutory mandate" (FOMC, 2016).

In addition to controlling inflation, central banks are also subject to other policy goals. For example, in the USA, the FOMC is "firmly committed to fulfilling its statutory mandate from the Congress of promoting maximum employment, stable prices, and moderate long term interest rates" and also states that "the median of FOMC participants' estimates of the longer run normal rate of unemployment was 4.9 per cent" (FOMC, 2016). The Bank of England must also "support the Government's economic objectives including those for growth and employment" (Bank of England, 2016). In Australia, the RBA must assist in the "the maintenance of full employment in Australia"; and more generally "the economic prosperity and welfare of the people of Australia" (Reserve Bank Act, 1959).

A third policy objective of moderating asset price growth is sometimes mentioned, but not formally acknowledged. Indeed, the US Fed, like some other central banks, prefers to avoid speaking in public about asset prices on the grounds that it is not mandated to consider them. Further, if a central bank is seen to be acting to constrain rising asset prices, this could generate political controversy, particularly among those with interests in seeing such prices rise.

Central banks in many developed economies set their interbank interest rate target each month, but its level is dictated by the judgment of the independent committee in the light of the prevailing policy priorities. Since the demand for reserves by the banks is relatively inelastic, accommodative behaviour by the central bank is typically not required following an announcement of a change in the target rate as the overnight rate will move quickly to the target.

Lender of last resort and financial stability

In a crisis, an important role played by the central bank is to operate as a **lender of last resort**, providing reserves on demand to financial institutions. Originally, this was to stop a bank run. When the Global Financial Crisis (GFC) impacted, banks could not refinance their positions in assets because their creditors demanded payment as liabilities matured. Central banks around the world had to step in to provide the refinancing and prevent the financial system from collapsing. Central banks have unlimited capacity to provide currency to the financial system on behalf of the government of the day. During the GFC, the Australian government also provided a deposit guarantee to give bank customers reassurance that their deposits were safe. Given that the US dollar is the major international reserve currency, the US Fed lent trillions of dollars to foreign banks and central banks during the GFC, essentially acting as a global lender of last resort to stop runs on banks outside the US that had issued liabilities denominated in the US dollar (see Felkerson, 2012).

23.4 Liquidity Management

Introduction

In this section we will build on the analysis from earlier chapters of liquidity management (undertaken by central banks to ensure that their target interbank rate is achieved), and we will also consider different institutional arrangements.

MMT shares the view that the central bank cannot control either the money supply or the level of bank reserves. Instead the central bank must accommodate the demand for reserves. Thus, the supply of reserves is best characterised as horizontal, at the central bank's target rate. That is the **endogenous money, horizontal reserve approach**, which was developed over the 1970s and 1980s by Moore and other post-Keynesians (see Lavoie, 1984; Moore, 1988). Most economists, regardless of their schools of thought, now accept this is a correct representation of the operating procedures of modern central banks.

However, the arguments for a horizontal supply of reserves provided by the central bank were formulated without considering:

- The operation of fiscal policy; and
- The situation when the central bank's target overnight interest rate is either near zero or equal to the interest rate paid to the banks on their reserve deposits.

When government spends, bank reserves increase initially since providers of goods and services to the government are paid by their deposit accounts being credited, which leads to the reserves of their banks being credited. When the interest rate paid by the central banks on reserves held by the banks (the support rate) is less than the target interbank rate, then the central bank would have to drain any excess reserves by selling government bonds (debt). Otherwise, the interest rate at which banks lend to and borrow from each other would fall below the target rate due to the operation of market forces. In the case of a fiscal surplus, where tax payments exceed government spending, bank reserves are net debited, so open market purchases may need to be conducted to augment the stock of bank reserves. In Chapter 20, we provided a simplified outline of the process of liquidity management by the central bank in the context of the implementation of fiscal policy. The essence of the argument that reserves are supplied elastically at the target rate is unchanged when fiscal policy is considered.

According to the logic of the **mainstream textbook money multiplier model**, the central bank could increase the money supply by injecting reserves through an open market purchase. This would enable greater bank lending, given the assumption that credit creation increases mechanically in line with the requirements of the **Fractional Reserve System** or the (predictable) ratio of deposits to reserves chosen by the banks (see Chapter 10 for a simple exposition of credit creation). However, this mainstream argument fails to recognise that the added reserves in excess of the banks' desired reserves would immediately drive the interbank rate to zero or to a non-zero support rate, since reserve requirements would not change until the following accounting period. If the target and support rates were unequal, the desire to achieve a positive target interbank rate would force the central bank to sell securities to drain those excess reserves that had just been added.

On the other hand, the mainstream view is that the central bank can reduce the money supply by taking reserves out of the system. But if there were no excess reserves to take out, then some banks would have insufficient reserves if such an action were undertaken. The central bank would have no choice but to add reserves back into the banking system to keep the market (interbank) rate at its target level.

In both cases, the level of bank reserves and the money supply would remain unchanged. Thus the central bank does not have the discretion to alter the supply of reserves to pursue an ostensible policy objective of controlling the money supply. Instead, an increase in the money supply, resulting from credit creation by banks will cause changes in the monetary base when banks adjust their reserves to bring them in line with a desired (or required) reserve-to-deposit ratio. The causation is not from a change in reserves (the monetary base) driving credit creation, but rather *from credit creation leading to a change in desired (or required) reserve holding* which is accommodated by the central bank.

Different interest rate setting arrangements

In the aftermath of the GFC, the USA and Japan adopted a near-zero interest rate target (see Table 23.1), so excess reserves could be left in the banking system. In the USA the target overnight interest rate target was between zero and 25 basis points (0.25 per cent) until December 2015. In that case, irrespective of the level of excess reserves held by the banks, the market rate remained within that range. In December 2015 the USA raised its target rate by 25 basis points to 25–50 basis points, which was followed by a further 25-basis-point increase in December 2016. The Fed began to pay interest on reserves at the low end of the target range, and to charge the rate at the high end of the target range on its loans of reserves. This ensured that the Fed funds rate (the overnight market rate) would generally stay within the target range, even if the Fed left excess reserves in the banks. In 2009 the Bank of England set both the bank (target) rate and the support rate at 0.5 per cent, so again excess reserves could be left in the banking system, since no bank would lend reserves to another at a lower rate than 0.5 per cent. This rate was cut to 0.25 per cent in August 2016.

Thus, in the circumstances of either a positive target rate equalling the support rate, or a near-zero target rate, the central bank can leave banks with excess reserves without compromising monetary policy, so that they are not forced to play an accommodative role. However, a shortage of reserves would drive the market rate above the target rate and would require a response from the central bank. In other words, payment of interest on reserves (or a decision to adopt a zero interest rate policy) produces an asymmetry: the central bank can leave excess reserves in the system, but it cannot leave the banks short of reserves.

Table 23.1 Target interbank rates for developed economies

Country	Interest rate (December 2016)	Previous change to rate
UK	0.25% from August 2016	0.5% from March 2009
USA	0.50% to 0.75% from December 2016	0.25% to 0.5% from December 2015
Japan	−0.10% to 0.00% from February 2016	0.00% to 0.10% from April 2010
Australia	1.50% from August 2016	1.75% from May 2016
ECB	0.00% from March 2016	0.05% from September 2014

Recognition that central banks can leave banks with excess reserves led to a belief that central banks might be able to use *quantitative* approaches to policy. In other words, perhaps central banks might regain some control over the money supply, at least in the upward direction, by filling banks full of excess reserves. This might then supplement what had come to be seen as the normal interest rate transmission mechanism of monetary policy. Given that interest rates were already at low levels after the GFC, could the central bank deliver still more stimulus by providing excess reserves?

The key question then is whether the capacity of central banks to exercise some discretion in respect of liquidity management means that the debate over whether reserves drive deposits, and hence the money supply, should be revisited. The answer is no, because a profitable loan requires a creditworthy borrower and an adequate interest rate differential between lending and borrowing rates for the bank to make an adequate return. The presence or absence of sufficient reserves has an incidental impact on this calculation.

The Bank of England has made major concessions about the role of banks in the financial system, in that it now rejects the simplistic intermediation role of banks using deposits as a basis for credit creation (McLeay et al., 2014). However, some central banks still appear to take the view that the monetary base, that is, the level of bank reserves plus the holdings of cash by the non-government sector, influences credit creation by the banks (this is called quantitative easing; see Section 23.6).

23.5 Implementation of Monetary Policy

Transmission mechanism

Here we examine how changes in monetary policy are alleged to impact on the macroeconomy through the operation of a transmission mechanism. Typically, the yield curve (YC) slopes upwards due to uncertainty about monetary policy and interest rate risk (see Chapter 10). Let us consider a cut in the target interbank rate. The yield curve is a graphical depiction of the **term structure of risk free interest rates** and plots the maturities of different government bonds on the horizontal axis against their respective yields (rates of return) on the vertical axis (see Figure 10.1).

Then the fall in the (target) interbank rate is likely to reduce medium- to long-term risk-free interest rates on public debt through arbitrage and hence the YC is likely to shift downwards. Arbitrage also plays out with respect to interest rates on private sector assets. Thus a change of interest rate policy is likely to affect private sector short-term rates (for example, on bank deposits) and also long-term rates on business and consumer loans. In Australia, these have normally been adjusted in line with a rise in the target rate, but only partially adjusted when the target rate has been cut. However if the change in the target rate is widely anticipated by the markets, it may be already factored in with respect to interest rates in general, so the movement of the YC after the announcement may be less clear cut.

Investment is a component of aggregate expenditure that is considered to be interest sensitive. New physical capital investment is usually based on borrowed funds. These projects are only undertaken if the expected net profit rate exceeds the interest rate on borrowed funds. Here the long-term interest rate is relevant, since the construction phase of major investment projects may last many months, if not years, and revenue from the project will not be generated until production starts and sales are made. All else being equal, more projects would be expected to make a sufficient return to justify additional spending following an interest rate cut. Long-run expectations of an uncertain future, which underpin profit expectations, are quite volatile, however, and reflect confidence. Thus, there is no guarantee that a 25 or 50 basis point cut in borrowing rates would have a positive impact on investment spending. In addition, the proposed level of investment say in 2019 will be the outcome of detailed planning in prior months and is unlikely to be sensitive to modest interest rate changes in the short term.

Durable consumption spending (for example, cars, houses and white goods, i.e. major home appliances) is often financed by borrowing, so borrowing costs, and hence the capacity to service a loan, are also relevant to these decisions about spending. In an uncertain economic climate, the degree of job security in the years ahead will also be an important consideration in terms of the capacity to repay the loan. A fall in the mortgage rate will reduce the periodic mortgage repayment, which would enable higher spending and/or increased saving, and the latter could be used to pay off the mortgage more quickly.

A fall in the target rate lowers domestic interest rates in general, and reduces the interest rate differential between the country in question and other countries. International capital flows (sales of financial assets to and purchases from foreigners) are claimed to respond to interest rate differentials. With a fall of the domestic rate, there is likely to be a fall in capital inflow (that is, lower sales of domestic assets to foreign investors relative to the purchase of foreign assets by domestic citizens) and so the local currency is likely to depreciate. This means that imports are more expensive when priced in the local currency, and are less attractive to buy, whereas exports are cheaper when their prices are denominated in a foreign currency, so net exports may increase, adding to any rise in total spending associated with higher investment and durable consumption expenditure.

While small changes in long-term interest rates (following corresponding changes in the target rate) may have little impact on spending, higher and higher long-term interest rates will eventually diminish domestic spending that is interest rate sensitive.

Thus, reliance on monetary policy to impact on aggregate expenditure, and indirectly on the inflationary process, is highly problematic. We briefly review the performance of the Reserve Bank of Australia with respect to inflation and unemployment outcomes in Section 23.7.

23.6 Unconventional Forms of Monetary Policy

Introduction

In developed countries adversely affected by the GFC, including the US, UK, Japan and the Eurozone, official rates were frequently cut in the early months of 2008, down to historically low levels. With the exception of the USA, which raised the federal funds rate to a range of 0.25 per cent to 0.5 per cent in December 2015 and a further 25 basis points in December 2016, these historically low rates persisted through to the end of 2016.

Policymaking since the advent of the GFC highlights the fact that monetary policy is still considered to be the main arm of macroeconomic policy. Since the GFC, inflation has been low for the developed economies and indeed the prospect of deflation has been often discussed, particularly in the Eurozone countries. The GFC rapidly manifested itself as a real crisis with the collapse of private sector spending. Fiscal stimulus measures were supported by the Organisation for Economic Cooperation and Development (OECD) and the International Monetary Fund (IMF), but with the proviso that such measures should be discontinued if adverse deficit/debt dynamics occurred. Stimulus measures were adopted in some countries, including Australia, the UK and the USA, in 2009 but were discontinued by 2010. Thus, there has been a reliance on monetary policy. However, with rates close to zero in the major developed economies, there has been limited scope for further cuts in official rates. Consequently, these economies have adopted unconventional forms of monetary policy.

Quantitative easing (QE)

In the absence of a capacity to reduce the intercept of the yield curve (YC), since their target interbank rates have been so low, central banks in Japan, UK and USA have resorted to trying to flatten the YC by so-called quantitative easing (QE). In simple terms, this means that central banks in these countries have developed programmes for buying longer-term treasury debt, and in some instances, private sector financial assets from both the bank and non-bank private sector. The objective here is to boost demand for these financial assets, which increases their market prices and thus lowers their yields, thereby flattening the YC. In addition, the impact of the private sector selling financial assets to the central bank is that bank reserves increase, which contribute to the overall growth of bank reserves in these countries.

As discussed previously in this chapter, the arrangements for setting interest rates, post-GFC, in the UK, Japan, the USA and the ECB, enabled QE to be implemented. With near-zero interest rates, and with a move to pay low interest rates on excess reserves, central banks were able to adopt an asymmetric policy: while they could not leave banks short of reserves, they were able to leave large quantities of excess reserves in the banking system, without adverse effects. If the central bank was forced to defend the target rate via bond sales, due to excess reserves, the initial sale of bonds would be reversed, thereby nullifying any macroeconomic effect.

The Bank of England identified the mechanisms by which QE is supposed to promote spending. First, purchases of financial assets reduce both short- and longer-term rates and promote greater confidence. This is defined as the signalling channel. Second, the purchases of long-term securities by the central bank gives rise to capital gains for households who hold these assets and they are assumed to consume part of their increase in wealth. This is referred to as the portfolio (re)balance channel. Third, the increased bank deposits and reserves created by asset purchases from non-bank institutions will increase the availability of bank credit, so banks may be more willing to lend (Joyce et al., 2012). This is the bank funding/lending channel.

MMT challenges whether the direct reduction of long-term rates via QE would be any more successful in stimulating these economies than the earlier cuts to official rates which indirectly reduced long-term rates. The main problem in many developed economies is sluggish growth and the consequent absence of large numbers of creditworthy firms seeking to borrow. Second, there may be some modest increases in spending from the wealth effects, but equally interest incomes for retirees are based on interest rates and will decline as rates fall. Third, MMT would reject the bank funding/lending channel since it is premised on reserves driving loans and deposits – in short, the money multiplier mechanism.

Negative interest rates

A second form of 'unconventional' monetary policy is the Swedish and Japanese central banks' resort to negative target interbank rates in February 2015 and late January 2016, respectively.

Japan cut its benchmark rate to −0.1 per cent. However, to reduce the impact on the earnings of financial institutions, the reserves held by financial institutions were divided into three tiers to which a positive, zero and negative interest rate were paid. In addition, the Bank of Japan also committed to continue its QE programme by ongoing purchases of Japanese government bonds (JGBs) worth about 80 trillion yen per year. These two measures were expected to reduce the intercept of the yield curve and "exert further downward pressure on interest rates across the entire yield curve" (Bank of Japan, 2016). The objective of this strategy was to "achieve the price stability target of 2 per cent at the earliest possible time".

It is curious to impose a new public tax on the private sector when the objective is to achieve a higher inflation rate, but this is a cautious move compared to ECB's imposition of a 0.3 per cent tax on all reserves, given a target rate of −0.3 per cent. Banks would have a greater incentive to economise on their holdings of reserves, but this does not mean that banks have an increased incentive to create credit.

Borio and Disyatat (2009: 19) note that:

> *A striking recent illustration of the tenuous link between excess reserves and bank lending is the experience during the Bank of Japan's "quantitative easing" policy in 2001–2006. Despite significant expansions in excess reserve balances, and the associated increase in base money, during the zero interest rate policy, lending in the Japanese banking system did not increase robustly.*

They went on to state that the reason for the slow credit growth at that time was that Japanese businesses had sufficient capital stock to satisfy the demands of a very weak consumption sector and did not need to borrow. There has been a recovery in the investment to GDP ratio since 2010, but the ratio remains well below the ratio prevailing in the early 1990s.

Conclusion

Consistent with neoliberal thinking, governments have continued to rely on monetary policy as the primary macroeconomic policy tool since the GFC. It has largely failed to address the higher unemployment in the developed economies. Consequently, given the limited scope for further conventional monetary easing, central banks have resorted to unconventional monetary policy. From a MMT perspective these measures would be ineffective. The evidence so far supports this view.

23.7 Monetary Policy in Practice

Figure 23.1 examines the performance of the Reserve Bank of Australia (RBA) with respect to inflation rate and unemployment rate outcomes (see Figure 23.1). In mid-1993 a two to three per cent inflation target for the RBA was introduced. There was a fundamental change in the relationship between inflation and unemployment following the 1991 recession, which was analysed in Chapter 18. From about 2000, the inflation rate exhibited a trend decline and this was accompanied by falling unemployment until about 2008. From late 2010, the unemployment rate rose steadily, albeit with some fluctuations, despite a series of reductions in the official (cash) rate, which culminated in the reduction from 1.75 per cent to 1.5 per cent in August 2016 (see Table 23.1).

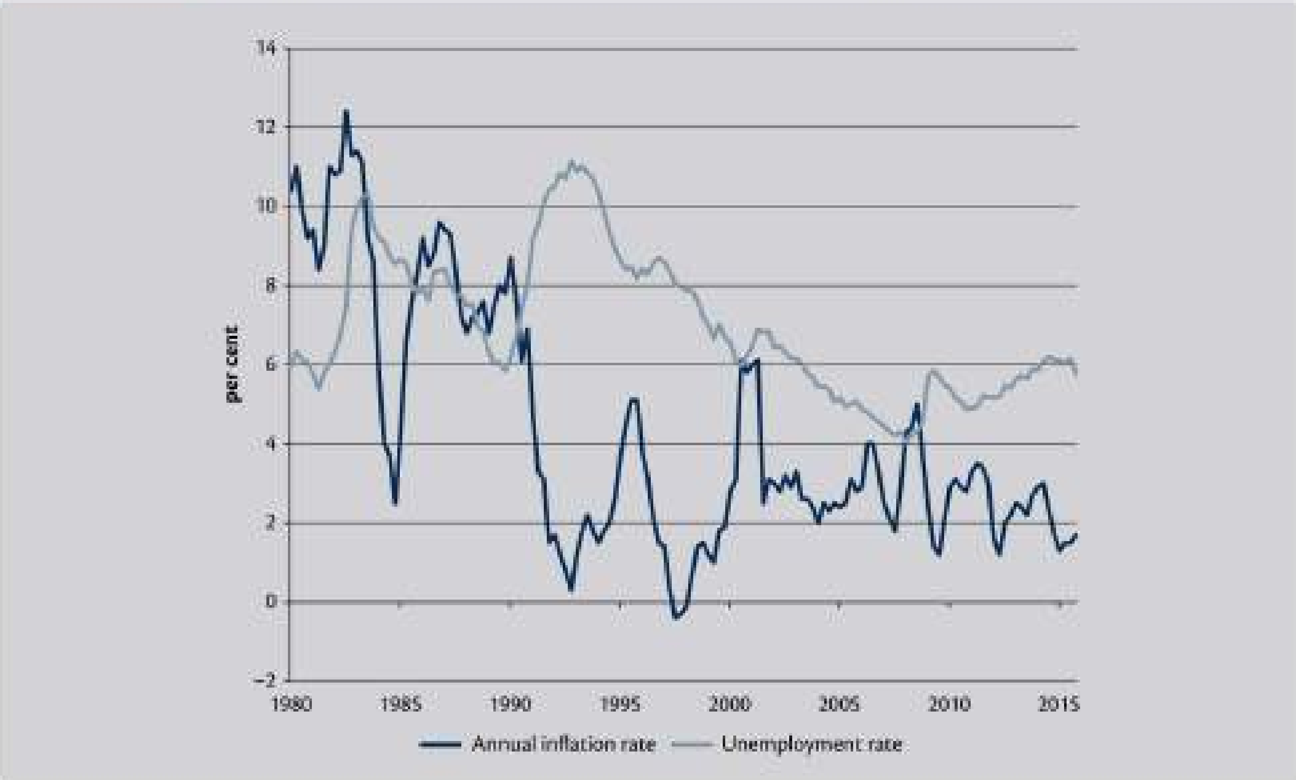
Thus, at a time when stimulus measures were justified, with the inflation rate below two per cent and unemployment rising, cuts to the cash rate have simply been ineffective, which is consistent with MMT. This issue is illustrated even more clearly when the performance of the peripheral Eurozone economies, such as Greece, Spain, Portugal and Ireland, is analysed. Despite persistent low inflation and periods of deflation (negative price level growth) and a low target interbank rate (0.00 per cent in 2018) set by the European Central Bank, these countries were mired in high unemployment in 2018; for example, Ireland with 5.9 per cent, Spain with 15.1 per cent, and Greece with 19.5 per cent (Eurostat, 2018).

23.8 The Advantages and Disadvantages of Monetary Policy

Monetarists and orthodox economists, in general, prefer to use monetary policy through interest rate setting as the main arm of macroeconomic policy because it is:

- Easy to implement (monthly) and flexible;
- Less subject to political interference; and
- More clearly understood by financial market traders.

Figure 23.1 Unemployment and inflation rates for Australia, 1980 to 2015, per cent



Source: Authors' own. Data from RBA Statistics (CPI All Groups quarterly, annualised) unemployed persons as percentage of labour force.

In addition, if the inflation rate is subject to a target level, say two per cent or a range of two to three per cent, then it provides an anchor for the expectations of consumers and business about future inflation.

The disadvantages of monetary policy include:

- A blunt, indiscriminate policy, which is not guaranteed to work in a timely manner, whether it be stimulatory or contractionary.
- An inappropriate policy response to, say, a cost push inflation, such as one driven by an external oil price shock from political uncertainty in the Middle East, or a widespread, severe drought.
- A single instrument, which in some circumstances, is attempting to affect three policy targets (inflation, GDP growth and asset price growth) which may be in conflict. Tinbergen (1952) noted that the number of instruments must equal the number of targets in order for economic policy to be consistent.
- Not regionally specific, so a housing boom in large cities may warrant a target rate increase to reduce the growth of house prices, via higher (variable interest) mortgage repayments, but at the same time falling house prices and declining job opportunities in rural areas may justify a rate cut.
- Often too tight if it is geared to a low inflation rate (or a target range), which can impose major economic and social costs of higher unemployment, which are often underestimated.

23.9 Central Bank Independence

Introduction

An important debate in the context of the institutional arrangements for the conduct of monetary and fiscal policy is the appropriateness of having an independent central bank within a modern monetary economy. If so, what should be the nature of the independence? We have already shown that the central bank must be accommodative in terms of its provision of reserves, unless either the target and support rates are equal or there is a near-zero target rate, but it is able to exercise some discretion in its setting of the target interbank rate, within the constraints of the policy objectives specified in legislation.

Rationale for independence

Some sovereign economies, such as Australia, New Zealand, Canada, and the UK, are not subject to legal restrictions on their central banks participating in the primary market for government debt. However in practice, central banks in those countries, with the exception of Canada, have limited engagement in the primary market. This prohibition was written into US law from the founding of the Federal Reserve Bank in 1913, although exceptions to the prohibition have at times been made, such as during the Second World War, when fiscal deficits reached 25 per cent of GDP.

Certainly, there is a consensus amongst the orthodox economics profession, as well as policymakers, that central bank financing of fiscal deficits by buying treasury bills on the primary market should be forbidden. The UK Treasury, via the Debt Management Office, chooses to fully fund its financing requirement by selling debt in line with a pre-announced schedule, which can be revised during the year. It provides the following rationale, which is representative of arguments made in support of this principle:

> *[T]he Government believes that the principles of transparency and predictability are best met by full funding of its financing requirement; and to avoid the perception that financial transactions of the public sector could affect monetary conditions, consistent with the institutional separation between monetary policy and debt management policy.* (HM Treasury, 2012: 8)

The first point implies that the tendency for there to be excessive government spending, over and above tax receipts (that is, a deficit) will be revealed by the issue and sale of debt in line with the overall financing requirement.[1] Under the auction system the quantities of debt of different maturities which are offered for sale will be sold with the yields determined by the market, so full financing will be achieved. Second, the sale of the debt will neutralise the impact on reserves that is associated with running a fiscal deficit. This may be thought to avoid the need for the central bank to maintain the integrity of monetary policy, by selling debt to soak up the excess reserves, which would then maintain the institutional separation between monetary policy and debt management policy. However, the banks may wish to hold some additional reserves and the non-bank private sector may wish to hold additional cash in view of the increase in economic activity. Thus the requirements of liquidity management may well mean that the additional debt held by the non-government sector is inconsistent with full funding of the financing requirement and there will not be the institutional separation of monetary policy and debt management policy.

Another case is in the Eurozone where the European Central Bank was presumed to be prohibited from buying the debt of the member governments. By design, these governments were supposed to be disciplined by markets to keep their deficits and debt within the criteria set by the Maastricht Treaty, that is, a three per cent deficit-to-GDP ratio and a 60 per cent debt-to-GDP ratio. These plans did not come to fruition, as we learnt earlier. The ECB's balance sheet is even bigger than that of the US Federal Reserve Bank, through the purchase of government debt of Eurozone members in the secondary markets.

Central banks in most developed economies are answerable to their parliament (or congress) and are obliged to provide detailed information about their operations and fiscal policy. As we noted earlier in this chapter, most parliaments specify macroeconomic objectives that guide central bank policy, which often include reference to low inflation, high (full) employment, acceptable growth, and financial stability. Parliaments may not be prescriptive about the instruments to be used to achieve these objectives, including whether to use open market operations or the discount window.

There is a commonly held view that the 'independence' enjoyed by many central banks enables them to be insulated from political pressures from special interest groups, despite the committees or boards which are responsible for setting interest rates often being appointed by the political party in power. Consequently, they are supposed to be able to make decisions that may not be popular, but have long-term economic benefits. Finally, despite some mainstream economists believing that the central bank should exercise some direct control over taxation and spending decisions, which are implemented by treasury, in practice this does not occur. Central banks do not refuse to make payments for treasury because they are focused on ensuring the smooth functioning of the payments system. If a central bank were ever to bounce a treasury cheque for 'insufficient funds', the head of the bank would be called before the elected representatives to provide an explanation. Sovereign government spending is constrained by fiscal commitments, not by central bankers.[2]

23.10 Horizontal and Vertical Operations: An Integration

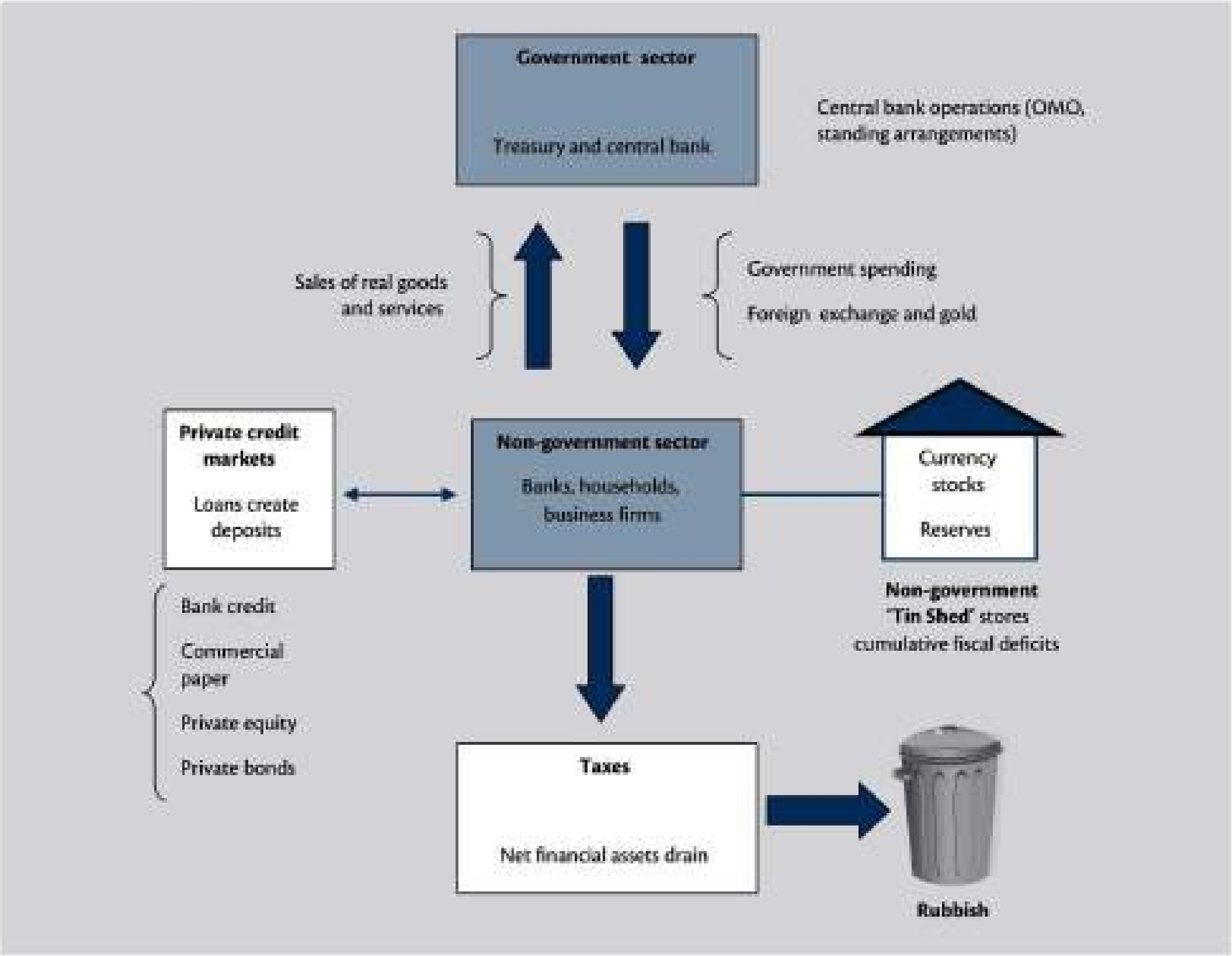
In some sense, the verticalists[3] and the horizontalists have each captured elements of the money supply process. One can conceive of a vertical component of the money supply process that consists of the government supply of fiat money; money drops vertically to the private sector from government through government purchases of goods and services as well as central bank purchases of assets (such as gold, foreign currency and bonds, and also through the discounting of assets held by banks).

Figure 23.2 depicts the vertical and horizontal aspects of the macroeconomic relations, the latter demonstrating the way private credit markets (for example, banks) participate in expanding and contracting the money supply.

Recall from our discussion above and in previous chapters that the private sector is willing to accept government fiat money, say from government spending, because the government imposes tax liabilities on the private sector which are payable in the fiat currency.

The government sector (treasury and central bank) injects 'currency' (broadly defined) into the economy. The treasury 'spends' currency while the central bank mostly 'lends' currency into existence. Tax payments (which discharge the liability) then drain fiat money, which can be pictured as a vertical movement from the private sector to government (and, hence, 'into the rubbish bin' as the money is simply wiped off the liability side of the central bank's balance sheet). The net difference between these vertical flows (deficit spending) leads to accumulation of fiat money hoards (currency in the hands of the public plus bank reserves), the 'tin shed', much of which is held by private institutions against the liabilities they issue. The government can also offer to exchange interest earning

Figure 23.2 Vertical and horizontal macroeconomic relations



bonds for non-interest earning cash and reserves. These bond sales – either sales by the treasury in the primary market, or by the central bank in secondary markets – also drain currency. What is left is held as cash deposits by households and firms and as reserves by banks.

On the other hand, we can think of the **bank money supply process as horizontal**; it can be visualised as a type of 'leveraging' of the hoarded vertical fiat money. Clearly, bank money is only one type of leveraging of the fiat money. A partial list of other types of leveraging would include commercial paper, private bonds, all types of bank liabilities, indeed all IOUs denominated in the fiat money of account. All of these private IOUs share three characteristics: they are denominated in the fiat money of account, they consist of long (owning a financial asset) and short (a future commitment to deliver a financial asset that is not currently owned) positions and they are 'inside' debt, such that the longs and shorts net to zero. A bank deposit can be thought of as a long position in fiat money, while the bank's borrowers have short positions, betting that they will be able to obtain money for repayment later.

A reduction of government spending can starve the non-government sector of funds so that those with outstanding loans to banks are not able to obtain sufficient money to make payments on loans. This is sometimes called a **short squeeze**. The borrowers, who might be workers, lose income because the cut in government spending causes unemployment to rise, or they might be business firms who experience a decline in revenue as sales fall and cannot obtain the money required through additional borrowing (an increase of the horizontal money supply). The squeeze arising from a cut in government spending could be eased if the 'savers' – those with positive bank deposit balances – were willing to increase their own spending or if others were willing to come into the market to take new short positions (lending to those squeezed). In that case, a funds shortage in the non-government sector could be relieved by operations in the horizontal section. But this is unlikely to occur when there is a slowdown in economic activity and the non-government sector adopts a cautious stance with respect to spending and lending.

The only reliable way a short squeeze can be relieved is if the government (via the vertical component) uses its capacity as the **net supplier of money**. If the government does not react to the short squeeze, the bank borrowers would be forced to try to sell assets, roll over loans or try to obtain new loans. This can lead to a fall in asset prices, which could degenerate into a general debt deflation. Defaults also occur with increased frequency if the short squeeze persists. On the other hand, this can be avoided if the central bank enters as lender of last resort (discounting assets or buying assets held by the private sector in exchange for cash) or if the treasury increases its deficits.

If things become sufficiently bad, banks become insolvent, with asset values below the value of liabilities as borrowers start to default on their outstanding loans. If the depositors with long positions 'liquidate' (demand fiat money instead of bank deposits), banks are forced to the discount window to borrow reserves. Beyond some point, as bank balance sheets deteriorate, they will not have sufficient capital (net worth) to obtain discount window loans from the central bank, requiring the deposit insurer to step in to 'resolve' the bank. As prices fall, borrowers default and banks fail, the private economy will almost certainly suffer a recession (or worse), lowering government tax receipts and perhaps raising government spending (through automatic stabilisers), which increases the government deficit (and available net saving).

Conclusion

In this chapter we have examined the MMT approach to monetary policy. We explained why central bankers realised that they are unable to control monetary aggregates and instead explicitly adopt an overnight interest rate target. In addition, we have discussed the role played by the central bank acting as the government's bank. This exposes the common belief that the central bank is "independent" of the treasury as a fallacy. The central bank must cooperate closely with the treasury to ensure the payments system functions smoothly.

We also examined what are called "unconventional" monetary operations that central bankers adopted in the aftermath of the GFC. We explained why these policies were largely ineffective in accomplishing what they were supposed to achieve. Finally, we have offered recommendations for more effective policymaking, proposals that take into account the understanding that MMT has brought to the field of monetary policy. The preferred policy would follow the recommendations of functional finance: the treasury ought to spend enough to maintain the full employment of labour. Central banks should provide the mix of currency (cash and reserves) and government bonds desired by banks, households and firms at the full employment level of output. There is no reason to pursue a 'short squeeze' below a position of full employment, although this is what the principles of 'sound finance' dictate. Unfortunately, policymakers in recent decades have been more inclined to pursue sound finance rather than functional finance.

References

Bank of England (2016) *Monetary Policy Framework*. Available at: http://www.bankofengland.co.uk/monetarypolicy/pages/framework/framework.aspx, accessed 13 February 2016.

Bank of Japan (2016) *Introduction of Quantitative and Qualitative Monetary Easing with a Negative Interest Rate*. Available at: https://www.boj.or.jp/en/announcements/release_2016/k160129a.pdf, accessed 21 February 2016.

Borio, C. and Disyatat, P. (2009) "Unconventional Monetary Policies: An Appraisal", Bank for International Settlements Working Paper no. 292.

Eurostat (2018) *Unemployment Statistics*. Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php/Unemployment_statistics accessed 22 September 2018.

Felkerson, J.A. (2012) "A Detailed Look at the Fed's Crisis Response by Funding Facility and Recipient", *Public Policy Brief No. 123*, Levy Economics Institute. Available at: http://www.levyinstitute.org/pubs/ppb_123.pdf, accessed 8 May 2016.

FOMC (Federal Open Market Committee) (2016) *Statement on Longer-Run Goals and Monetary Policy Strategy*. Adopted effective 24 January 2012; as amended effective 26 January 2016.

HM Treasury (2012) *Debt and Reserves Management Report 2012–13*, March 2012. Available at: https://www.dmo.gov.uk/media/2061/drmr1213.pdf, accessed 17 September 2018.

Joyce, M., Miles, D., Scott, A. and Vayanos, D. (2012) "Quantitative Easing and Unconventional Monetary Policy – An Introduction" *The Economic Journal*, 122, F271–88.

Lavoie, M. (1984) "The Endogenous Flow of Credit and the Post Keynesian Theory of Money", *Journal of Economic Issues*, 18(3), 771–97.

McLeay, M., Radia, A. and Thomas, R. (2014) "Money Creation in the Modern Economy" *Bank of England Quarterly Bulletin*, 54 (1), 14–27.

Moore, B. (1988) *Horizontalists and Verticalists: The Macroeconomics of Credit Money*, Cambridge: Cambridge University Press.

Reserve Bank Act (1959) Commonwealth of Australia. Available at: https://www.legislation.gov.au/Series/C1959A00004, accessed 8 May 2017.

Tinbergen, J. (1952) *On the Theory of Economic Policy*, Amsterdam: North Holland.

Endnotes

1. The deficit could be 'bad' in that it is the outcome of automatic stabilisers.
2. However, the 'debt limit' in the US is a restriction that has sometimes come close to forcing the government to stop making planned fiscal payments. When the total outstanding stock of federal government debt nears the debt limit, Congress must raise the limit, which is normally done routinely. However, on occasions political maneuvering can lead to an impasse when a block of representatives will refuse to vote to raise the limit (usually with a demand that some programme they do not like is cut from the fiscal strategy). When this happens, treasury must try to postpone issuing debt, sometimes by postponing payments, and prepare for a freeze on government spending. The consequences can be severe, including defaulting on promised interest payments on outstanding debt. This is a not a constraint imposed by the Federal Reserve Bank. Congress could amend the law to eliminate the debt limit. Alternatively, Congress could amend the rules governing *how* Treasury makes payments to reduce the routine issue of treasury bonds (see Chapter 20 for discussion of US operating procedures). Crises occurred in 2011 and 2013 due to an impasse over raising the debt limit. In early February, 2018, President Trump signed a **bill suspending the debt ceiling** until 1 March, 2019.
3. 'Verticalists' is usually used to describe orthodox economists who conceive of the money supply curve as vertical in money, interest rate space. The idea is that the central bank determines the money supply exogenously (through the deposit multiplier process); see Chapter 10. The discussion here, however, uses the term 'verticalist' in a different sense.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

24 POLICY IN AN OPEN ECONOMY: EXCHANGE RATES, BALANCE OF PAYMENTS AND COMPETITIVENESS

Chapter Outline

Learning Objectives

- Understand the components of the balance of payments and their interrelationship.
- Acknowledge the distinction between the nominal and real exchange rates.
- Analyse the role of trade in the determination of equilibrium national income.

24.1 Introduction

In Chapter 15, we introduced trade into the income/expenditure model. The exposition was simplistic in the sense that we assumed that exports are determined by the income levels prevailing in the rest of the world (that is, they were exogenous to the domestic economy) and that imports are a simple proportion of the national income of the home economy. This proportion was termed the **marginal propensity to import**.

In this chapter, we extend our understanding of the way in which the economy behaves once it becomes open to the world. We will continue to consider the price level to be fixed, which means we are assuming that firms respond to an increase in aggregate demand by increasing real output. Later in the chapter we will consider price level movements in the open economy context.

We will consider the concept of an exchange rate and examine how movements in exchange rates influence exports and imports and financial transactions between nations.

For an economy as a whole, imports are goods and services coming into the nation from abroad and as such represent a real benefit to residents. Conversely, exports are goods and services that are sold in foreign markets.

Exports represent a real cost to residents because they represent resources (labour, capital and other productive inputs) that the residents are unable to utilise to produce goods and services for their own use. It is obvious that the only motivation for a nation to export, and incur the real costs involved in exporting goods and services abroad, is to gain foreign currencies, which in turn allow the nation to purchase other goods and services that it does not produce itself.

If imports exceed exports then a nation is able to enjoy a higher material living standard by consuming more goods and services than it produces for foreign consumption. We will consider how this conception of trade interacts with a flexible exchange rate.

You will already appreciate that transactions between nations involve both goods and services and financial flows. The financial transactions represent currency flows into and out of a nation and have significant implications for movements in the exchange rate and other macroeconomic aggregates, such as interest rates, the inflation rate and real GDP.

All transactions between a nation and the rest of the world are recorded in the balance of payments. We will initially examine the way the national statistician accounts for the external economy via the balance of payments, which is a framework that is closely related to the national accounts.

24.2 The Balance of Payments

Residents (households, firms, and governments) of every nation conduct economic transactions with residents of other nations and all these transactions are recorded in the international accounts for each nation. The international accounts are made up of a number of component accounts:

- The international investment position (IIP) which "shows at a point in time the value of: financial assets of residents of an economy that are claims on non-residents or are gold bullion held as reserve assets; and the liabilities of residents of an economy to non-residents" (IMF, 2011: 7).
- The balance of payments which is "a statistical statement that summarises transactions between residents and non-residents during a period. It consists of the goods and services account, the primary income account, the secondary income account, the capital account, and the financial account" (IMF, 2011: 7).
- All "other changes in financial assets and liabilities accounts" (valuation changes and so on) (IMF, 2011: 7).

The Balance of Payments and related accounts are compiled by national statistical agencies (such as the UK Office of National Statistics, the Australian Bureau of Statistics (ABS), the US Bureau of Economic Analysis) using an international standard set down in the International Monetary Fund's *Balance of Payments and International Investment Position Manual (BPM6)* (IMF, 2011), augmented by the System of National Accounts (SNA) (United Nations Statistical Commission, 2009). While there are variations in the terminology used by different nations, the principles are universal.

The IMF manual is the "standard framework for statistics on the transactions and positions between an economy and the rest of the world" (IMF, 2011: 1). The differentiating feature of these various accounts relates to "the nature of the economic resources provided and received" by the nation (IMF, 2011: 9).

Like any accounting framework, the balance of payments is based on a double entry debit and credit system of record. Every transaction that is recorded has two equal and offsetting entries, each of which corresponds to the inflow and outflow of funds. Here we just record the net flows.

Credit entries consist of transactions where foreign residents make payments to local residents. Examples include exports of goods and services, income receivable from investments abroad, reductions in external assets, or increases in external liabilities.

Debit entries consist of transactions where local residents have to make payments to foreign residents. Examples include imports of goods and services, income payable abroad, increases in external assets, or a decrease in external liabilities.

Balance of payment examples

Example 1: Export of goods and services

An Australian resident sells $A1,000-worth of goods to a resident in the US.

The Balance of Payments in Australia will record:

Credit: Exports $A1,000

Debit: Currency $A1,000, increase in financial assets

Example 2: Australian resident borrows from a US bank

An Australian resident takes out a loan for $A1,000 from a US bank.

The Balance of Payments in Australia will record:

Credit: Loan $A1,000, increase in liabilities

Debit: Currency A$1,000, increase in financial assets

Table 24.1 shows how the ABS presents the balance of payments data for Australia and the USA. Observe the heading structure: Current Account and Capital and Financial Account being the major subaccounts of the Balance of Payments. Then within each of the subaccounts are several other italicised subheadings, which record different elements of the transactions between Australia and the rest of the world.

We will now briefly discuss the Current Account and the Capital and Financial Account with reference to Table 24.1.

The current account

The current account records all current transactions between a nation's residents and non-residents in goods and services, primary income and secondary income (see row 1 in Table 24.1).

Table 24.1 Australian and US balance of payments, current prices

		2016–17 $A(m)	2017 $US(m)
1	Current Account	−29,469	−449,141
2	*Goods and Services*	12,165	−552,277
3	Goods	13,819	−807,495
4	Services	−1,654	255,219
5	*Primary income*	−39,811	221,732
6	*Secondary income*	−1,823	−118,596
7	Capital and Financial Account	24,611	356,605
8	*Capital Account*	−762	24,746
9	Acquisition/disposal of non-produced, non-financial assets	−182	
10	Capital transfers	−580	24,746
11	*Financial Account*	25,373	331,859
12	Direct investment	70,827	−24,393
13	Portfolio investment	−1,945	212,487
14	Financial derivatives	−711	−23,074
15	Other investment	−23,168	165,149
16	Reserve assets	−19,629	1,690
17	Net errors and omissions	4,858	92,536

Source: Data from ABS (2017); US Bureau of Economic Analysis (2017).

The goods and services or balance of trade records "transactions in items that are outcomes of production activities" (IMF, 2011: 149) and reflect exchanges between the local economy and the rest of the world (row 2). The data are typically obtained from information collected from exporters and importers by the nation's customs department.

Exports and imports of goods relate to movable or tangible goods (row 3), while services are all products other than tangible goods (row 4). Services include items such as banking and insurance, transport, and education. While items bought by tourists while on holiday may be tangible, all such expenditure is recorded as services under the IMF conventions used.

Primary income "represents the return that accrues to institutional units for their contribution to the production process or for the provision of financial assets and renting natural resources to other institutional units." (IMF, 2011: 183)

There are two categories of primary income (row 5):

- Income that is associated with the production process, for example, wages paid, taxes and subsidies on production. If a resident is paid for labour by a non-resident then primary income is deemed to have been earned.
- Income that is associated with the ownership of financial assets, for example, dividends and interest.

These flows are accounted for in the primary account if they are current. You will appreciate that they impact on the measure of national income in the national accounts.

The secondary income account (row 6) relates to current transfers between residents and non-residents. These do not add to national income, but rather involve redistributions of income between nations. There is nothing of economic value that is exchanged in return for a secondary income transfer. Typical secondary income account transactions include personal transfers (remittances to or from overseas), charitable contributions, social benefits (such as pension payments to or from abroad), and current taxes on income and wealth.

Economists are often focused on the current account because the transactions it records are of direct relevance to the determination of national income. Our earlier discussions about the sectoral balances and the income/expenditure determination all explicitly considered the current account of the balance of payments.

Exports (injections) and imports (drains) are key components of aggregate demand.

The capital account and financial account

While the current account of a nation tends to focus on transactions with the rest of the world, which impact on the measurement of national output and income, the capital account (row 8) is the financial side of these transactions.

What would happen if a nation exported more than it imported? Ignoring the primary and secondary accounts for the moment, the net outflow of goods and services would be accompanied by accumulating financial claims against the rest of the world. This is because the demand for the nation's currency to meet the payments necessary for the exports would exceed the supply of the currency to the foreign exchange market to facilitate the import expenditure.

How might this imbalance be resolved? There are several possible ways. A most obvious solution would be for foreigners to issue liabilities to the domestic residents. This would lead to a net accumulation of foreign claims (assets) held by residents. This item would be recorded in the capital account as a debit because it enhances the capacity of non-residents to make transactions in the local economy.

Another solution would be for non-residents to draw down local bank balances, which means that net liabilities to non-residents would be lower.

As you can see from rows 8 to 10 in Table 24.1, the capital account records the transactions covering non-produced non-financial assets and capital transfers between residents and foreigners.

The financial account (rows 11 to 16) is a balancing account, recording the "net acquisition and disposal of financial assets and liabilities" (IMF, 2011: 10).

24.3 Essential Concepts

Before we consider a more complex income and expenditure model that incorporates exchange rates, we need to understand the basic nomenclature.

The following essential concepts are used in open economy macroeconomics:

- Nominal exchange rates.
- Foreign exchange markets.
- Exchange rate determination mechanisms: fixed and flexible.
- Real or effective exchange rates, unit labour costs and competitiveness.

We will consider the history of exchange rate systems in a later section of this chapter.

Nominal exchange rate (e)

The nominal exchange rate (e) is the number of units of one currency that can be purchased with one unit of another currency. We can quote a bilateral exchange rate in two different ways. Consider the relationship between the Australian dollar ($A) and the United States dollar ($US).

We might be interested in knowing the amount of Australian currency that is necessary to purchase one unit of the US currency (one US dollar). In this case, the $US is what we call the **reference currency** and the other currency is expressed in terms of how much of it is required to buy one unit of the reference currency. So $A1.25 = $US1 means that it takes $1.25 of Australian currency to buy one $US.

Alternatively, e can be defined as the number of US dollars that are needed to buy one unit of Australian currency (one Australian dollar). In this case, the $A is the reference currency. So, using the previous example, this is written as $US0.80= $A1. Thus, if it takes $A1.25 to buy one $US, then $US0.80 is required to buy one $A.

The quotation under the first alternative with the US dollar as the reference currency is the inverse of the second alternative. But to understand exchange rate quotations you must know which currency is the reference currency.

In this chapter, we use the convention that e is the amount of foreign currency which is required to buy one unit of the domestic currency. Thus:

e is the amount of $US which is required to buy one unit of the domestic currency ($A).

Change in the nominal exchange rate, appreciation and depreciation

Imagine that an Australian resident wishes to buy a product from a US supplier who quotes the current US price as $US36, and the $US to $A parity is currently at $0.80. Then the equivalent Australian price is $A45 (divide the foreign price by the nominal exchange rate). This situation is shown in the first row of Table 24.2.

What happens if the nominal exchange rate falls to 0.60 (as shown in the second row of Table 24.2)? This means that instead of 80 cents US being required to purchase one $A only 60 cents US is required.

So, **a fall in e means that the $A has depreciated**, and one Australian dollar (the reference currency) is worth less in terms of foreign currency. In the example shown in Table 24.2, this would mean that the price of the product from the USA would now be equal to $A60 ($US36 divided by 0.60). Thus, even though the quoted US dollar price for the product remains unchanged, the local price equivalent is now higher when the nominal exchange rate depreciates.

Table 24.2 Comparison of international prices

e $US to $A1	Foreign price $US	Local price equivalent ($A)	Change in $A compared to starting value
0.80	36	45	–
0.60	36	60	Depreciation
1.00	36	36	Appreciation

The example shows that a depreciation of the $A leads to:

- Foreign goods being more expensive in terms of their $A price, and, other things being equal, this should lead to a **fall in the quantity of imports demanded**.
- The price that foreigners must pay for Australian goods being lower for a given $A price. Other things being equal, this should lead to a **rise in the quantity of exports demanded**.

Now, assume that the USA to Australian parity rises from 0.80 to $1.00. This means that we now need one US dollar to purchase one Australian dollar. So, given our exchange rate definition, **a rise in *e* means that the $A has appreciated**.

In the example shown in Table 24.2, this would mean that the price of the product from US would now be equal to $A36 ($US36 divided by one). The example shows that an appreciation of the $A leads to:

- Cheaper foreign goods in terms of their $A price, and other things being equal, this should lead to a **rise in the quantity of imports demanded**.
- Higher prices which foreigners will have to pay for Australian goods for a given $A price for Australian-produced goods. Other things being equal, this should lead to a **fall in the quantity of exports demanded**.

What determines the exchange rate?

Sometimes we refer to **foreign exchange** in jargon as **forex**. The supply of and demand for currencies are in turn linked to trade and capital flows between countries, and also to relative interest rates and expected changes of interest rates between countries. The determination of exchange rates is exceedingly complex, and movements of rates are impossible to predict. No theory or model of exchange rates has been capable so far of predicting exchange rate movements. Views held by economists diverge on whether the transactions associated with trade dominate those relating to cross-border transactions of financial assets or vice versa.

Exchange rates are thought to be influenced by the supply of and the demand for currencies in the world foreign exchange markets, which could be the local bank foreign currency desk or elsewhere, like a train station kiosk in a city where travellers meet. However, saying that supply and demand influence exchange rates really tells us very little about the real world determination of exchange rates.

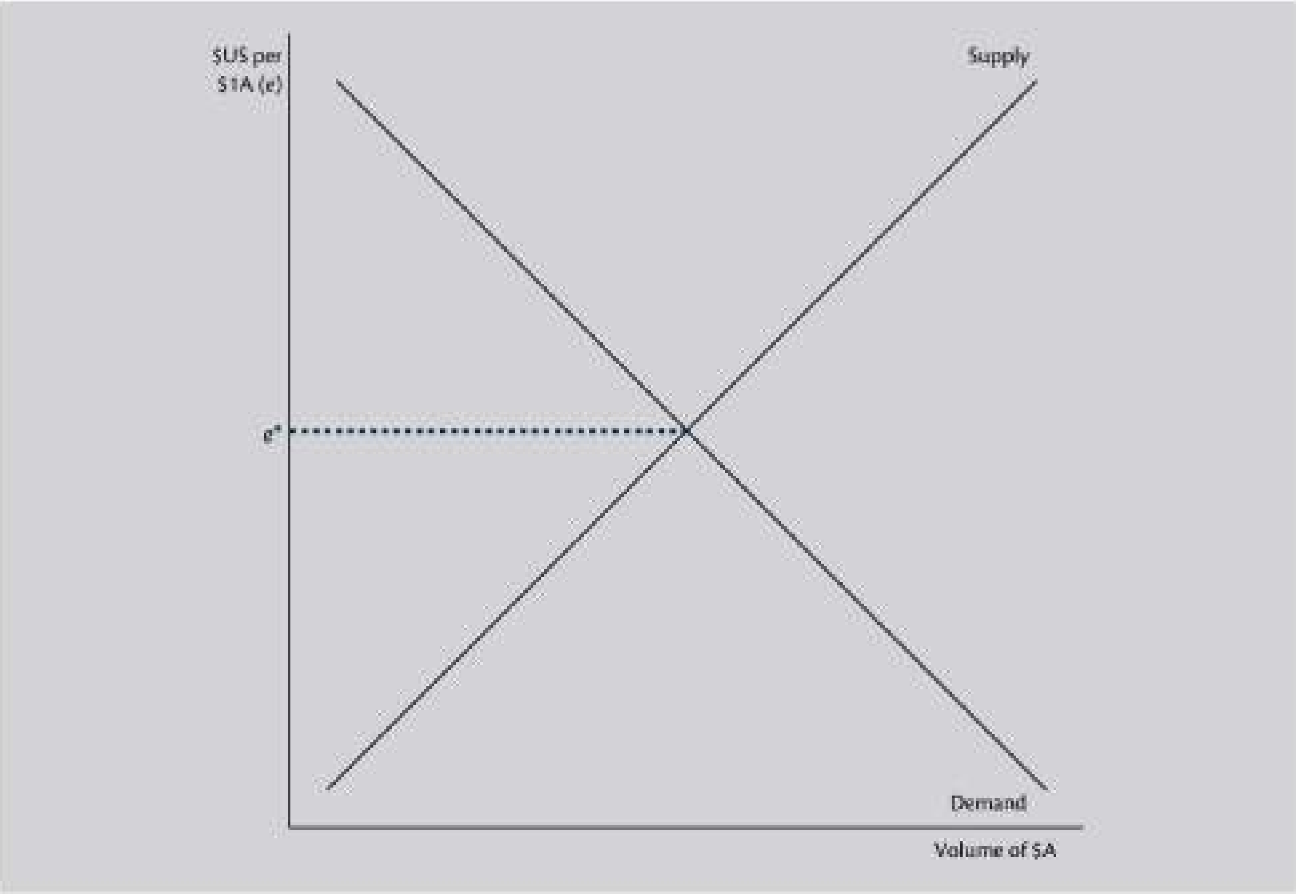
Economists often present a very simple model of exchange rate determination that is based on the usual supply and demand curves. The quantity of the currency is placed on the horizontal axis while the 'price' shown on the vertical axis is expressed as the exchange rate. The student should keep in mind, however, that this is a highly simplified way of looking at exchange rate determination.

In Figure 24.1 we consider the orthodox approach to the foreign exchange market for the $A and the $US. Most currencies are traded in the foreign exchange market, but we have just selected two. We will focus only on exports and imports in analysing the supply of and demand for $As.

Consider the supply of Australian dollars to the foreign exchange market. When Australian residents buy foreign goods (imports), buy foreign assets or lend abroad, they need to purchase the foreign currency in which the transaction is denominated. To buy the currency they desire, they supply $As in exchange. We focus on imports. Here the assumption is that, as the exchange rate (*e*) increases, more $A are supplied, so the supply curve of $A is upward sloping. Assuming a fixed import price in the foreign currency, the $A price of imports has fallen, which increases the quantity demanded. What happens to the total $A value of imports when the exchange rate appreciates depends upon something economists term the **price elasticity of demand**. Elasticity is the responsiveness in percentage terms of demand to price changes.

- When demand rises less in percentage terms than the price falls, we consider the demand for the commodity to be inelastic. Total revenue (or spending) will fall.
- When demand rises more in percentage terms than the price falls, the demand for the commodity is said to be elastic. Total revenue (or spending) will rise.
- When price and quantity change by the same proportion the demand for the commodity has a unitary elasticity and total revenue (or spending) does not change.

Figure 24.1 A simple bilateral foreign exchange market



The upward-sloping supply of $A in Figure 24.1 would be explained by the demand for US imports being elastic, so that a lower $A price of US imports leads to more $A being spent on US imports.

Alternatively, on the demand side, when foreigners buy Australian goods and services (exports) and/or Australian financial assets, they require $A. They purchase them in the forex market by supplying their own currency in exchange. We focus on exports. The downward-sloping demand for $A can be explained by the US demand for Australian exports being elastic.

- In the orthodox approach to exchange rate determination, e is in equilibrium when supply equals demand.
- If there is an excess demand for $A (that is, demand exceeds supply) then there is pressure for the $A to appreciate in price relative to other currencies. As noted, an appreciation means that one unit of the reference currency ($A) buys more US dollars, that is e rises.
- If there is an excess supply of $A (that is, supply exceeds demand), the $A depreciates and one unit of the reference currency ($A) buys less US dollars, so e falls.

These changes in e resolve the supply and demand imbalance (see Figure 24.1). In the case of a depreciation in the Australian dollar, the foreign price of Australian exports is now lower (less $US required to purchase a given $A-priced good), and with an elastic demand for exports which varies inversely with price (by assumption), the demand for exports rises. This translates into an increase in the demand for $As. On the other hand, with an elastic import demand, the supply of $As falls following the depreciation of the $A against the $US.

We have shown that if both the Australian demand for imports and the foreign demand for Australian exports are elastic, then the trade balance improves following a depreciation. However, these conditions are overly demanding (i.e. sufficient but not necessary).[1]

The circumstances under which the trade balance unambiguously improves following a depreciation is referred to as the **Marshall-Lerner Condition**. It states that **net exports will improve following a depreciation**

as long as the sum of the export and import price elasticities exceeds unity. You do not have to learn the proof underpinning this condition.

In summary, if the Marshall-Lerner condition is satisfied:

- An excess supply of $A in the foreign exchange market leads to a depreciation (*e* falls) and a rise in net exports. This will reduce the excess supply of $A in the foreign exchange market.
- An excess demand for $A in the foreign exchange market leads to an appreciation (*e* rises) and a decline in net exports. This will reduce the excess demand for $A in the foreign exchange market.

Another component of the current account is net income, which results from the foreign ownership of domestic assets and vice versa. This pattern of ownership of assets gives rise to a net flow of dividend and interest payments. If the net flow is positive, then national income rises, other things being equal. If the net flow is negative, then national income falls, other things being equal.

The net income flows of the Australian current account are negative. In this case, a depreciation in the $A can lead to improved net income if the interest payments or dividends are denominated in a foreign currency. The gain through this part of the current account would supplement any gains that are made as a result of the impact of the depreciation on the trade balance.

For simplicity, we shall ignore the possible impact on net income and assume that, through the satisfaction of the Marshall-Lerner condition, a depreciation of the domestic currency not only improves the trade balance, but also the current account balance.

We can define three trade balance outcomes:

- The trade balance is in deficit if the local currency value of its exports is less than the local currency value of its import spending.
- The trade balance is in surplus if the local currency value of its exports is greater than the local currency value of its import spending.
- The trade balance is in balance if the local currency value of its exports is equal to the local currency value of its import spending.

Take Australia, as an example. A trade deficit means that increasing quantities of Australian dollars are being accumulated by non-Australian residents. In return, the non-Australian residents have supplied goods and services (imports) to Australian residents.

Clearly, the foreigners have allowed Australia to run a trade deficit because they preferred to accumulate financial assets denominated in Australian dollars. The alternative would have been to spend the Australian dollars they acquired through their exports to buy Australian goods and services (that is, to buy Australian imports).

Had the foreigners used their entire export income, which is denominated in $A, to purchase other goods and services from Australia, then there would have been a trade balance.

A trade deficit thus means that the foreigners are increasing their nominal savings (which in this case manifests as Australian dollar-denominated financial assets).

Finally, while most currencies float freely against each other, at times the central bank of a country will enter the foreign exchange markets as a buyer or seller of its (local) currency as a means of influencing the parity determined in that market. This is called **official intervention**.

It must be remembered that the simple supply and demand approach presented in this section really cannot explain exchange rate determination in the real world. The most important flaw in this approach is the focus on international trade in goods and services. In reality, financial transactions are many orders of magnitude greater, at least for the world's major currencies.

The size of the financial market explains why a country like the US can run persistent trade deficits, and a country like Japan can run persistent trade surpluses, without there being persistent pressure on their exchange rates. At best, supply and demand analysis is helpful only for what are called 'spot markets' such as the sale and purchase of bananas at a farmers' market at the end of the last market day of the week. The 'market' for foreign exchange is much more complicated, and is influenced by expectations of future exchange rates as well as interest rates.

The demand for a country's financial assets will play a big role in determining exchange rates. Most of the demand for say, the Australian dollar is not for the currency *per se* but rather for Australian dollar-denominated financial and real assets. Likewise, most of the global demand for the US dollar is not for purchases of US goods and services, but rather for financial assets denominated in $US that can be held as earning assets in diversified portfolios.

It is important to note that prices in asset markets are affected by interest rates. In foreign exchange markets, participants are concerned with *relative* interest rates (that is, domestic interest rates relative to foreign interest rates) and *future* exchange rates. The decision to hold $US assets will thus be affected by current US interest rates relative to foreign interest rates, but also by expectations about the appreciation or depreciation of the $US relative to other currencies.

This leads us to two rival theories concerning exchange rate determination. The orthodox view, which focuses on trade in goods and services, proposes the *purchasing power parity theorem*. According to this view, exchange rates will tend to move to equalise exchange rate adjusted prices. In other words, whether one buys a McDonald's Big Mac in the US or in Australia, one should pay the same price, adjusted for exchange rates. If the burger costs four dollars in the US, and if the exchange rate is $US 0.80 per $AU, then the price in Australia should be five Australian dollars.[2] At this point, the *purchasing power* of the two currencies would be equal (at 'parity'). Obviously, the price will deviate from this if there are differential transactions costs, including transportation costs. The important point is that the purchasing power parity approach supposes that it is transactions in goods and services that drive the exchange rate.

The alternative approach follows Keynes' theory, which focuses on asset markets. It is called the *interest rate parity theorem*. This proposes that exchange rates are in equilibrium when expected returns to asset holders are equalised across exchange rates.[3] In other words, if a portfolio holder is choosing between an Australian government bond and a US government bond, prices will tend to move to equalise total expected returns. The total return will include not only the expected interest rate paid but also the expected movement of the relative exchange rates. Let us say that the Australian bond pays five per cent but the US bond pays only three per cent. Why would anyone hold the US bond? They will do so only if they believe that the US dollar will appreciate sufficiently relative to the Australian dollar. In that case, the expected return on the US bond could equal the expected total return on the Australian bond, which includes the losses due to the depreciation of the $A for US bondholders, when they convert their $A back to $US. Note that this approach focuses on transactions involving financial assets rather than trade in goods and services.

For our purposes, it is important to keep in mind that the simplistic demand and supply approach presented in Figure 24.1 is useful for understanding how the orthodox economist goes about explaining exchange rate determination, but it is at best a partial theory. Indeed, it does not perform very well in explaining actual exchange rates. By contrast, the interest rate parity theorem performs relatively well across currencies.

International competitiveness

In the previous section, we learned that an appreciation(depreciation) of a nation's exchange rate leads to foreign goods becoming cheaper(dearer) in terms of the local currency, which should lead to a rise(fall) in the quantity of imports demanded, other things being equal.

Further, an appreciation(depreciation) of a nation's exchange rate means that foreigners have to pay higher(lower) prices in their currency for locally produced goods, which should lead to a fall(rise) in the quantity of exports demanded, other things being equal.

These conclusions, however, only focus on one element of the competitiveness of a nation's goods and services in international trade, the nominal exchange rate, e.

But to really answer the question of whether local goods and services are becoming more or less price competitive with respect to goods and services produced overseas, we have to relax the 'other things equal' assumption and consider the domestic and foreign inflation rates.

This leads us to define a new concept, the **real exchange rate** that depends on two factors:

- Movements in the nominal exchange rate, e; and
- Relative inflation rates (domestic and foreign).

There are also non-price dimensions to competitiveness, including quality and reliability of supply, which are assumed to be constant at this stage.

We define the ratio of domestic prices (P) to those in the rest of the world (P_w) as P_w/P. We call this the relative price because it expresses the foreign price level relative to the domestic price level. We assumed that P_w/P was constant when we analysed movements in the nominal exchange rate in the previous section.

If the nominal exchange rate (e) is fixed, then we can conclude that:

- If P_w is rising faster than P, then local goods are becoming relatively cheaper than foreign goods; and
- If P_w is rising slower than P, then local goods are becoming relatively more expensive than foreign goods.

The inverse of the relative price ratio (P/P_w) measures the ratio of export prices to import prices and is known as the **terms of trade**.

The real exchange rate

Movements in the nominal exchange rate and/or the relative price level provide information about movements in the relative trading competitiveness between nations. The real exchange rate measures the combined impact of these variables and is used to measure a nation's competitiveness in international trade because it adjusts the nominal exchange rate for the relative price levels.

The real exchange rate (R) is defined as:

$$(24.1) \qquad R = (P_w/e)/P = (P_w/Pe)$$

where P is the domestic price level specified in local currency (say, \$A), and P_w is the foreign price level specified in foreign currency units (say \$US).

The real exchange rate is the ratio of prices of goods abroad measured in \$A (P_w/e) to the \$A prices of goods at home (P). P_w/e takes the foreign price expressed in foreign currency units and converts it into an equivalent Australian dollar price at the current exchange rate. P_w/e and P are then in like units, so they can be compared directly and movements in the real exchange ratio (R) are unambiguous.

To better understand the relevance of the real exchange rate, consider the following example. Assume that P = \$A12, P_w = \$US10, and e = 0.8. Remember that a quotation of e = 0.8 means that it takes \$US0.80 to purchase one unit of Australian currency (that is, \$A1). Substituting for P, P_w and e in Equation (24.1), the real exchange rate is given by R = (\$US10/0.8)/12 = 1.042. Thus, US-produced goods are more expensive than those in Australia by a ratio of 1.042, or 4.2 per cent higher.

A rise in the real exchange rate can occur if:

- The nominal e depreciates; and/or
- P_w rises more than P, other things equal.

We consider a rise in the real exchange rate to signal that a nation has increased its international trade competitiveness and this should lead to an increase in local exports and a reduction in local imports.

A fall in the real exchange rate can occur if:

- The nominal e appreciates; and/or
- P_w rises less than P, other things equal.

We consider a fall in the real exchange rate to signal that a nation's international trade competitiveness has fallen and this should lead to a fall in local exports and a rise in local imports.

In Chapter 16, we considered the factors that might impact on the price level of a nation. If prices are set on unit labour costs, then the way to decrease the price level relative to the rest of the world is to reduce unit labour costs faster than everywhere else, or compress profit margins (mark-ups).

If the rate of growth of nominal wages is faster than labour productivity growth, then unit labour costs rise and vice versa. As we saw in Chapter 16, the real wage is a composite of the nominal wage determined in the labour market as a result of bargains between workers and employers, and the price level, which is determined by firms in the goods and services market.

The problem is that if a nation attempts to improve its international competitiveness by cutting nominal wages to reduce real wages, and in turn, unit labour costs, it not only undermines aggregate demand, but also may damage its productivity performance.

If, for example, workforce morale falls because of cuts to nominal wages, it is likely that industrial sabotage ('slacking' on the job) and absenteeism will rise, undermining labour productivity.

Further, overall business investment is likely to fall in response to the extended period of recession and wage cuts, which erodes future productivity growth. Thus, there is no guarantee that this sort of strategy will lead to a significant fall in unit labour costs, and if it were to be successful, there are likely to be adverse consequences for aggregate demand.

There is robust research evidence to support the notion that by paying high wages and offering workers secure employment, firms reap the benefits of higher productivity which yields improvements in a country's international competitiveness.

24.4 Aggregate Demand and the External Sector Revisited

In Chapter 15 we developed the income/expenditure framework to explain the factors that influence the production of real GDP (or real national income). That chapter used the national accounting concept of GDP at constant, rather than current prices, as our measure of economic activity.

In our income/expenditure framework in Chapter 15 we expressed the flow of total expenditure in any period as the sum of the following sources of spending:

- Consumption by households or persons (C)
- Investment spending by firms (I)
- Government spending (G)
- Export spending by foreigners (X) minus import spending by domestic residents (M), which we denote as net exports, $NX = (X - M)$.

From Chapter 15, we know that the equilibrium level of real national income (Y) is determined by aggregate demand (as long as prices remain unchanged), such that:

$$Y = E = C + I + G + (X - M) \tag{24.2}$$

In this section, we will develop a more detailed account of net exports, $(X - M)$, to take into account the influence of the real exchange rate and international competitiveness, which we discussed in the previous section.

In Chapter 15, our treatment of the determinants of net exports was deliberately very simple. We assumed that exports are given in any particular period and determined by national income in the rest of the world, which is beyond the influence of the local economy in question. We also assumed that a nation imports a fixed proportion of every dollar of national income. We called that proportion the **marginal propensity to import** (m) and defined it as the extra import spending that occurs following a dollar increase in national income.

In the previous section, we learned that movements in the real exchange rate, which is a summary measure of international competitiveness, influences net exports. We established that the higher is the value of the real exchange rate, the cheaper are locally produced goods and services to foreign buyers, which means that foreigners will purchase more of them. In other words, exports rise when the real exchange rate rises.

Further, the higher is the value of the real exchange rate, the more expensive are foreign-produced goods and services for local buyers, which means they will purchase less of them. In other words, imports fall when the real exchange rate rises.

There are many other factors in the real world that determine the demand for a nation's exports and the demand by residents for imports, which we abstract from here in order to focus on the most significant determinants.

We also abstract from adjustment responses, which are common in international trade. So a rise in the real exchange rate might only influence future exports on the expiry of existing export contracts, which tend to be multi-year. In the following analysis we are simplifying by assuming that the response between movements in the real exchange rate and changes in the flows of export and import spending occurs within the current period.

Let us consider exports. We now assume that the level of exports in any period is determined by the real exchange rate (R) and world income (Y_w) and we write this in the following way:

$$(24.3) \quad X = \lambda Y_w + \theta_X R$$

This might appear at first inspection to be daunting, but if you apply the techniques that we developed in Chapter 7, you will grasp the meaning of this equation fairly easily.

In Equation (24.3) the Greek letters next to world income (Y_w) and the real exchange rate (R) measure how responsive export spending is to changes in these two variables. The coefficient, λ measures how much a nation's export income rises as a result of a rise in world income. If you think about it from the perspective of the rest of the world, λ is its marginal propensity to import.

Similarly, the coefficient θ_X measures the responsiveness of exports to changes in the real exchange rate. Remember we are simplifying by assuming that this response is immediate in the current period.

From our theoretical exegesis, we conjecture that when either world income and/or the real exchange rate rises, exports will increase. Thus, we specify λ and θ_X to be positive.

Turning to imports, we assume that the level of imports that a nation purchases depends on both real national income (Y) and the real exchange rate (R). Thus:

$$(24.4) \quad M = mY - \theta_M R$$

The coefficient m is the marginal propensity to import and we know its value lies between zero and one. We conjecture that the impact of the real exchange rate will be negative. This means that when the real exchange rate rises, the nation becomes more competitive and foreign goods become more expensive in local currency terms, so that import spending falls. Then with the coefficient θ_M assumed to be positive, the responsiveness of imports to changes in the real exchange rate, takes the form $-\theta_M R$.

If we assume that domestic and foreign price levels are both constant, then changes in the nominal exchange rate (e) are solely responsible for movements in the real exchange rate, R. A higher(lower) nominal exchange rate causes a lower(higher) real exchange rate.

Thus, net exports (NX) depends on local real GDP, world real GDP and the real exchange rate (see Equation 24.8), where the real exchange rate impacts on both exports and imports.

24.5 Trade in Goods and Services, Product Market Equilibrium and the Trade Balance

National income equilibrium with trade

In this section, we continue to assume that P_w/P is constant, which means that domestic and foreign firms respond to increases in real aggregate demand by increasing real output rather than prices.

Spending on domestic goods determines real output and income. Total spending on domestically produced goods and services is equal to total spending by domestic residents minus their spending on imports plus foreign demand for exports.

Referring back to Chapter 15, we have the following behavioural equations, which comprise our theory of aggregate demand:

(24.5)	Consumption function	$C = C_0 + cY_d = C_0 + c(1 - t)Y$
(24.6)	Investment function	$I = I_0 - bi$
(24.7)	Government spending	G
(24.8)	Net exports	$NX = \lambda Y_w - mY + \theta R$

where $\theta = \theta_X + \theta_M$. Note that θ is the net impact of changes in the real exchange rate, here expressed as the nominal exchange rate because we assume that P_w/P is constant.

We can substitute the individual behavioural equations into the equilibrium income equation (24.2) such that:

$$(24.9)\quad Y = E = C_0 + c(1 - t)Y + I_0 - bi + G + \lambda Y_w - mY + \theta R$$

which, if we refer back to the way we simplified Equation (15.15a) to get Equation (15.15c), we can write as:

$$(24.10)\quad Y = (1/[1 - c(1 - t) + m]) \times [C_0 + I_0 - bi + G + \lambda Y_w + \theta R]$$

This expression for equilibrium national income tells us that real GDP (Y) will be the sum of all the expenditure terms that do not directly depend on national income $[C_0 + I_0 - bi + G + \lambda Y_w + \theta R]$ multiplied by $(1/[1 - c(1 - t) + m])$.

We can use Equation (24.10) to study what happens to national income when one of the terms in the square brackets changes.

The net exports function

Equation (24.8) represents net exports (NX) in terms of world income (Y_w), domestic national income (Y) and the real exchange rate (R). The real exchange rate is the net impact on aggregate spending of a real exchange rate change, taking into account the impacts on the individual components, namely exports and imports.

Figure 24.2 shows the net exports function expressed in terms of national income, so we assume that both the shift variables (Y_w and R) are constant. At low national income, there would be an overall trade surplus because the component of imports that is driven by domestic income (mY) is low and the impact of the real exchange rate R on net exports is positive via θR.

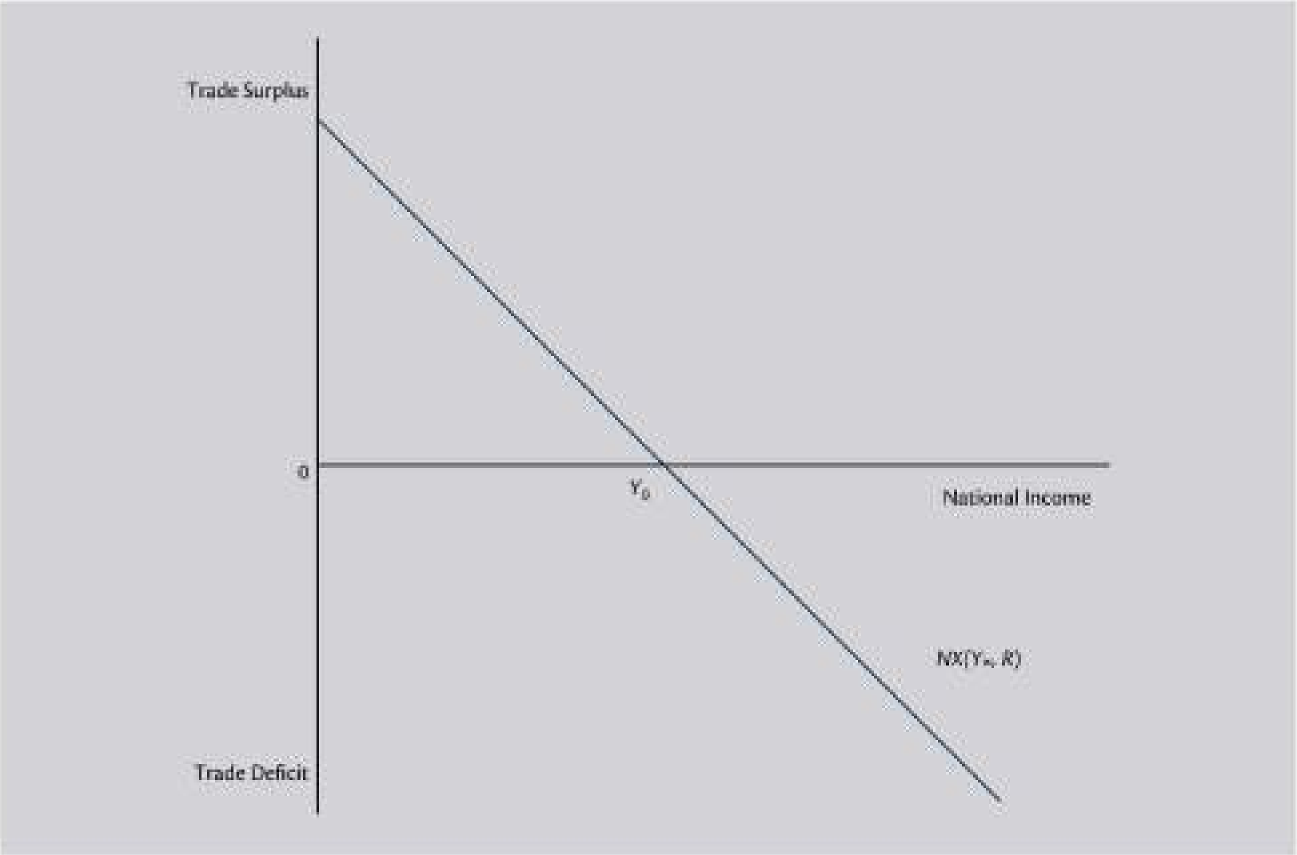
Then as national income rises, the trade balance will move from a surplus to a deficit. This is because as domestic national income rises, imports rise via the marginal propensity to import (m). Thus the net exports function has a negative slope. At point Y_{tf}, the net exports are balanced (exports equals imports).

To summarise, to the left of $NX = 0$ at Y_{tf}, there is a trade surplus because, for a given level of exports, the lower level of income leads to a smaller expenditure on imports. To the right of Y_{tf}, there is a trade deficit, because at the higher level of income, imports are higher relative to the fixed level of exports.

As noted, the net exports function is drawn against national income with the other variables in Equation (24.8), Y_w and R, being held constant. If either Y_w or R changed, the NX function would shift.

We have assumed that $\lambda > 0$ so that an increase in world income levels boosts our exports and so net exports rises (other things being equal). Further, in our discussion we assumed that the net impact on net exports of a change in the real exchange rate (θ) is positive. This means that a depreciation of the real exchange rate, that is a rise in R, improves the trade balance.

Figure 24.2 Net exports as a function of constant price national income



We conclude that:

- If world income rises(falls) the net exports function would move out(in) due to a change in the intercept, with an unchanged slope.
- If the real exchange rate depreciates(appreciates) the net exports function would move out(in), due to a change in the intercept, with an unchanged slope.

The slope of the net exports function is determined by the marginal propensity to import, m. The larger is the marginal propensity to import, the steeper the slope because the greater is the leakage from the expenditure stream per extra dollar of national income generated and the more quickly the trade balance moves into deficit at each income level.

We can use this understanding and that provided by the equilibrium national income expression (Equation 24.10) to study what happens to national income when world income and/or the real exchange rate change.

The impact on national income and net exports of a change in world income

We have seen (from Equation 24.10) that the equilibrium level of national income (and real GDP) depends on the level of domestic autonomous expenditure ($C_0 + I_0 + G$); the interest rate sensitive component of investment (bi); the level of world income (Y_w); and the real exchange rate.

What happens to national income if either the level of world income or the real exchange rate changes? In the analysis that follows we assume that the central bank supplies the monetary base at a constant discount rate, so that the interest rate charged by the commercial banks can be treated as constant.

In Figure 24.2 the trade balance line is denoted as $NX = 0$. With the other determinants constant (world income and the real exchange rate), we noted that there was one level of national income where the trade account would be balanced (where imports equal exports). We denoted that level of national income as Y_0 in Figure 24.2.

REMINDER BOX

We reintroduce the aggregate demand function from Chapter 15 (see Figure 24.3). Recall Figure 15.4 which showed that a rise in autonomous spending would lead to the aggregate demand function shifting up in parallel fashion (the shift in the intercept being measured by the change in autonomous spending).

We can generalise that insight into the current context by noting that the autonomous expenditure components are just the right-hand terms in Equation (24.8) which interact with the spending multiplier.

We can translate that knowledge into Figure 24.3 by drawing an $NX_0 = 0$ line which corresponds to national income level Y_0. All national income levels below Y_0 will result in trade surpluses because imports will be lower than exports, other things being equal. All national income levels above Y_0 will generate trade deficits, because imports will be higher than exports, other things being equal.

In our discussion, we also noted that the NX line would shift up if world income increased and down if world income decreased. This means that the $NX = 0$ line in Figure 24.3 will shift to the right if world income rises and to the left if world income falls.

The reason is simple. Start from national income level Y_0, which initially coincides with a trade balance $NX_0 = 0$. If world income rises then at that national income level (Y_0), exports will be higher than before and so the trade account would be in surplus. Trade balance would require a boost to imports which, in turn, would occur at higher levels of national income. We denote the level of national income corresponding to the new trade balance $NX_1 = 0$ as Y_1.

But we also know that a rise in world income will lead to the rest of the world importing more goods and services from the domestic economy, which means that exports rise. As we learned in Chapter 15, if any of the expenditure components rise, then the aggregate demand function shifts upwards.

Figure 24.3 denotes an initial point Y_0. Aggregate demand is at E_0. Our conclusions are not affected by the assumption that trade is balanced initially. The rise in exports pushes the aggregate demand function upwards to E_1 and the new national income equilibrium occurs at Y^*.

At this point, real GDP is higher, national income is higher and we could show that employment would be higher and unemployment lower once the cyclical labour supply adjustments that we studied earlier were exhausted. You will also note that the economy now has an external surplus, being to the left of the $NX_1 = 0$ line.

An increase in world income leads to a rise in net exports

We now confirm the claim about the impact of an increase in world income on net exports.

Prior to the rise in world income, the level of real GDP, Y_0 is consistent with a balance of trade, $NX_0 = 0$.

$$(24.11) \quad X_0 = M_0 + mY_0$$

where M_0 is a constant quantity of imports at the current constant exchange rate.

If we denote $\Delta X = \lambda\Delta Y_w$, which is the increase in exports resulting from the rise in world income, then national income Y_1 at which net exports are zero satisfies:

$$(24.12) \quad X_0 + \Delta X = M_0 + mY_1$$

If we subtract Equation (24.11) from (24.12) we get:

$$(24.13) \quad \Delta X = mY_1 - mY_0 = m\Delta Y^*$$

So the change in the level of real GDP, such that net exports are again zero is given by:

$$(24.14) \quad \Delta Y^* = \Delta X/m$$

On the other hand, the increase in equilibrium real GDP (national income) resulting from an autonomous increase in exports (driven by the rise in world income) is given by (from Equation 24.10):

$$(24.15) \quad \Delta Y = 1/[1 - c(1 - t) + m] * \Delta X$$

This magnitude is less than $\Delta X/m$ because $[1 - c(1 - t) + m] > m$. Thus, at the new national equilibrium, Y^*, net exports are positive, that is a trade surplus.

By way of summary, a rise in world income induces a rise in foreign purchases of the economy's exports in the same way as the local economy's import demand will be stimulated if its national income rises.

This has three effects:

- The aggregate demand line (E_0) shifts upwards by the initial injection of aggregate demand from exports (ΔX), giving rise to a new level of equilibrium national income Y^*. The increase in equilibrium national income is given by $1/[1 - c(1 - t) + m] \times \Delta X$, given the constant interest rate. This is just a specific example of the general result that the aggregate demand schedule shifts in response to changes in autonomous spending.
- The $NX = 0$ line shifts to the right from $NX_0 = 0$ to $NX_1 = 0$.
- Imports rise too because the shift in the aggregate demand line (to E_1) means that income levels are higher. But the shift in the $NX = 0$ line is greater than the increase in equilibrium income. Thus net exports rise, but by less than the rise in exports. This result holds even if domestic interest rates rise.

You should be able to work out what would happen if there were a world recession and world income levels fell. In this case, the local economy would experience a drop in aggregate demand because exports will be lower than previously and national income will fall.

In this case, the $NX = 0$ line also shifts to the left and if the economy was, for simplicity, initially experiencing balanced trade, then the new equilibrium income level will be associated with a trade deficit (the fall in exports is greater than the fall in imports).

24.6 Capital Controls

The history of financial crises indicates that large-scale financial speculation can undermine a nation's real economy relatively quickly if the government attempts to peg its currency to another currency or the economy has significant foreign currency-denominated debt exposure (private or public).

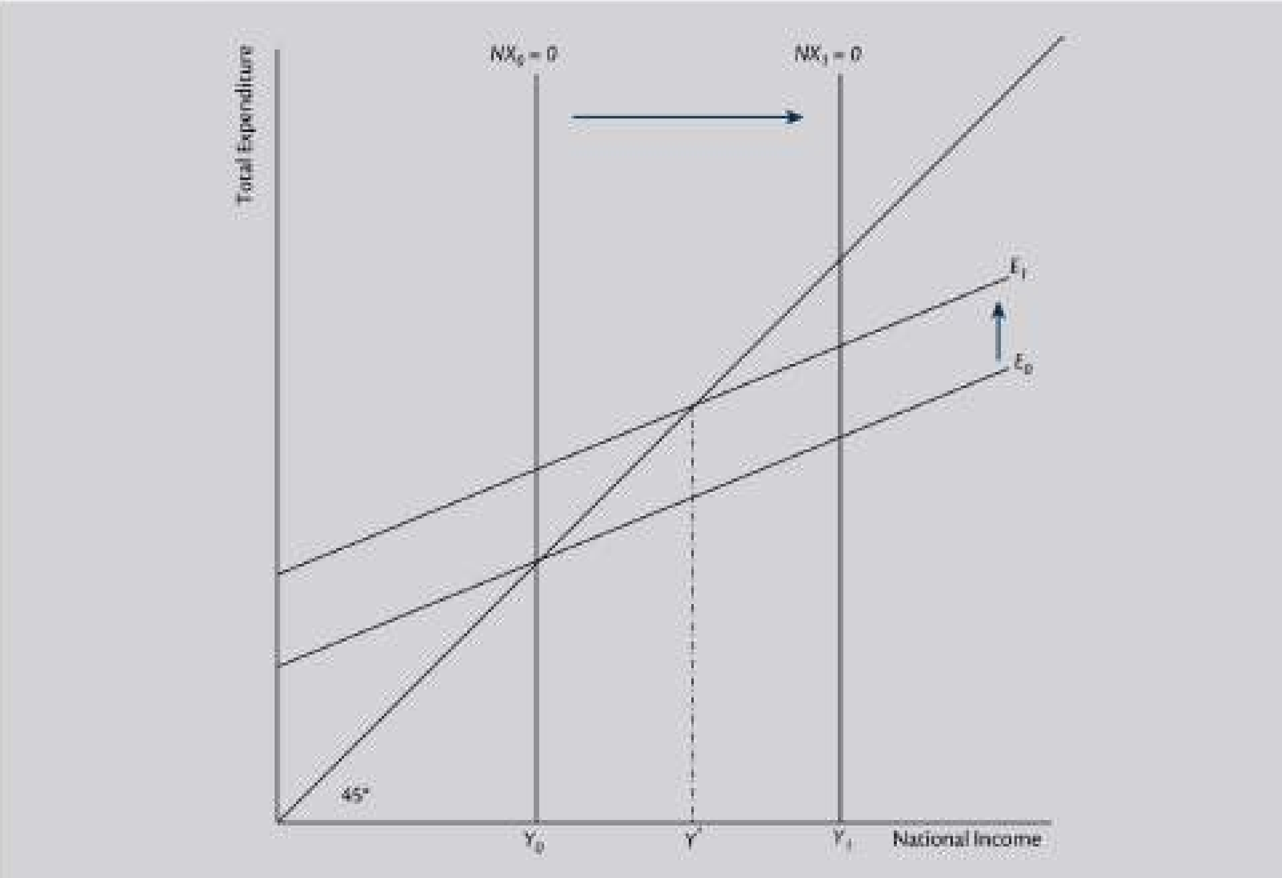
While the international community could agree that certain forms of speculative activity would be considered illegal, in lieu of that, the nation under attack must defend its own prosperity. One such suggestion is to introduce capital controls to limit the size and flexibility of international financial flows.

Capital controls are policies that restrict the free movement of capital, either in terms of inflows or outflows. There are broadly two types of capital controls:

- Administrative or direct controls, which impose limits or bans on capital flows.
- Market-based controls, which impose extra costs on capital flows to reduce the incentives to shift funds across national borders.

A government might, for example, place limits on foreign exchange transactions, international bank transactions, or bank withdrawals. Restrictions on movements of precious metals such as gold might also be considered.

Figure 24.3 Equilibrium national income with a change in world income



The aim is to limit the scope of speculative flows (in or out) to manipulate the exchange rate and strain the central bank's foreign exchange reserves.

Capital controls allow the central bank to run an autonomous monetary policy and the treasury to use fiscal policy to manage domestic demand in the interests of the nation.

The case for the use of capital controls is strongest in two circumstances: countries that desire to manage their exchange rates, and nations pursuing a development strategy. As we have seen, countries that do not float their exchange rates are subject to speculative attacks. While floating rates preserve the most domestic policy space and at the same time remove the incentive to speculate against the currency, some countries persist in either fixing their exchange rates, or in carefully managing them (sometimes within a narrow corridor). China is an example of such an exchange rate regime. In this case, the nation can guard against speculative attacks by imposing capital controls that make it difficult to exchange the currency. Again, China has historically tightly managed the flow of capital.

A country that is pursuing a development strategy in the context of a floating exchange rate might face the prospect of strong currency appreciation. Foreigners sensing profitable opportunities in the nation might rush in with short-term investment strategies to make quick profits. However, rising exchange rates can work against the development strategy because foreign currency prices of the nation's output rise relative to world prices. The developing nation might protect its new industries by constraining short-term capital flows to keep speculators from excessively appreciating the currency.

In September 1998, during the Asian debt crisis, the Malaysian government introduced capital controls after the currency had appreciated significantly and the central bank had pushed interest rates up towards 18 per cent and undermined the viability of many local businesses.

The case for the use of capital controls by a rich, developed country on a floating exchange rate is generally weaker. However, in the absence of effective international regulation of financial practices, even such a country might need to protect itself from abusive practices by foreign financial institutions.

Conclusion

This chapter examined macroeconomic policy in the open economy. The concept of the balance of payments was introduced. Essential concepts such as exchange rates, appreciation and depreciation of the currency, the real exchange rate, and international competitiveness were defined and discussed. We examined trade balances and discussed some of the factors that influence the balance of trade as well as the balance of payments. The Marshall-Lerner conditions were identified, helping to clarify the circumstances in which a currency depreciation(appreciation) might increase(decrease) a trade surplus. Effects of aggregate demand on the trade balance were also discussed. Finally, the topic of capital controls was introduced, with a distinction made between administrative and market-based controls.

References

ABS (Australian Bureau of Statistics) (2017) *Balance of Payments and International Investment Position, June, 2016*, Cat No. 5302, Canberra.

IMF (International Monetary Fund) (2011) *Balance of Payments and International Investment Position Manual (BPM6)*, Washington. Available at: http://www.imf.org/external/pubs/ft/bop/2007/pdf/bpm6.pdf, accessed 18 September 2018.

United Nations Statistical Commission (2009) *System of National Accounts 2008*, New York: European Communities, IMF, OECD, UN and World Bank.

US Bureau of Economic Analysis (2017) *International Data*, Table 1.1. U.S. International Transactions.

Endnotes

1. The elasticity condition for exports is overly demanding because even a unit price elasticity of demand for Australian exports will lead to an unchanged level of expenditure (in $USs) by Americans, which translates to a higher demand for $As given the depreciation of the $A. On the other hand, when considering the impact of a depreciation on the $A expenditure on imports, the depreciation only impacts through the change in the $A price and the $A expenditure on imports.
2. *The Economist* publishes a 'Big Mac' index which is used as an informal test of purchasing power parity (PPP). In practice, the exchange rate-adjusted Big Mac price varies considerably around the world. Note that PPP should be applied to a *basket* of traded goods and services, not to a single commodity. The 'law of one price' is more restrictive than PPP, for it presumes that the exchange rate-adjusted prices of individual traded commodities (such as the Big Mac) ought to converge to equality across currencies. Note that PPP assumes free international movement of goods and services.
3. More formally, the interest rate differential between two currencies equals the differential between the forward exchange rate and the spot exchange rate. Note that interest rate parity assumes free international movement of capital.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

PART F

ECONOMIC INSTABILITY

CHAPTER

25 THE ROLE OF INVESTMENT IN PROFIT GENERATION

Chapter Outline

Learning Objectives

- Recognise that private sector investment is inherently unstable.
- Understand introductory theories of investment behaviour.
- Comprehend the macroeconomic relationship between investment and profits and specifically the direction of causation.
- Appreciate the significance of fluctuations in investment for the operation of the business cycle.

25.1 Investment in a Capitalist Monetary Economy

In Chapter 4 *National Income and Product Accounts*, we learned that investment spending, or gross fixed capital expenditure, is a major component of aggregate demand. In Chapter 6 *Sectoral Accounting*, we noted that investment forms one row in the transactions matrix, as does the term profits. In this chapter, we tie those two aspects of business activity together.

Profits and investment spending are intrinsically linked in a capitalist economy, which we described in Chapter 2. In this chapter, we examine the behaviour of private investment spending and explain where profits come from in a macroeconomic sense.

We prioritise a study of investment for three main reasons.

- Fluctuations in investment are a key driver of the business cycle, that is, economic activity, and hence employment. The other major components of aggregate demand are relatively stable.
- Investment spending is an important determinant of total profits in the economy.
- Investment spending can be significantly altered by government policy, which means that policy can 'manage' the economic cycle somewhat through its influence on private investment growth.

Economists use the term 'investment' differently to its common usage, which might include people placing their saving in the form of financial assets, real estate and other speculative vehicles.

In the financial papers and reports you will often read or hear, for example, that 'investors have become more pessimistic' in relation to bond prices or other financial assets. Use of the term 'investors' in this context is different to the way we use it in macroeconomics.

Investment is defined in macroeconomics to be a **flow of spending which is devoted to increasing or maintaining the stock of productive capital**.

The capital stock comprises factories, machines, offices, and other durable products that are used up in the production process, and also includes inventories and residential housing.

The volatility of investment

Table 25.1 shows total investment expressed as a ratio of GDP for selected countries from 2000, with decade averages for the 1980s, 1990s and 2000s. The coefficient of variation is defined as the ratio of the standard deviation to the mean and provides a normalised measure of dispersion which allows us to compare dispersion across very different samples without recourse to information about the underlying average.

Investment ratios have been falling in Germany, Japan and the UK and rising in the other countries. In Germany, Japan and the US, the investment ratio has become more volatile since the 1980s.

Table 25.1 Total investment ratios for selected countries, per cent of GDP

Year	Australia	France	Germany	Japan	UK	US
2000	25.0	22.4	23.9	25.1	18.8	23.6
2001	23.5	22.1	22.3	24.3	18.3	22.1
2002	25.2	21.2	19.9	22.5	18.2	21.6
2003	26.8	21.1	19.7	22.4	17.7	21.7
2004	27.2	21.8	19.1	22.5	17.5	22.5
2005	27.9	22.4	18.8	22.5	17.6	23.2
2006	27.3	23.2	19.8	22.7	18.0	23.3
2007	28.7	24.1	25.8	22.9	18.5	22.4
2008	28.9	24.1	25.9	23.0	17.4	25.8
2009	27.7	21.3	18.1	19.7	14.6	17.5
2010	27.1	21.9	19.6	19.8	16.0	18.4
2011	27.5	23.2	21.1	25.2	15.8	18.5
2012	29.1	22.6	19.3	25.9	16.0	19.4
2013	27.6	22.3	19.5	21.2	16.4	19.8
2014	26.8	22.5	19.8	21.8	17.3	25.0
2015	26.1	22.4	19.2	22.0	17.2	25.3

(Continued)

Table 25.1 Total investment ratios for selected countries, per cent of GDP (*Continues*)

Year	Australia	France	Germany	Japan	UK	US
Average						
1980s	26.8	25.9	24.5	29.6	18.5	25.5
1990s	24.6	18.9	22.6	28.8	17.4	18.7
2000s	27.0	22.4	20.7	22.7	17.2	21.9
Coefficient of variation						
1980s	7.3	6.4	6.0	5.9	10.5	4.9
1990s	5.7	10.2	6.1	8.5	7.0	6.5
2000s	5.3	3.9	6.9	6.5	6.5	8.7

Source: Data from IMF World Economic Outlook Database, October 2016.

Note: Both total investment and GDP are expressed in current local currency. Investment or gross capital formation is measured by the total value of the gross fixed capital formation and changes in inventories and acquisitions less disposals of valuables for a unit or sector.

Gross and net investment

Gross investment is defined as the total addition to the existing capital stock.[1] However, in any given period, some of the existing capital will become inoperative due to wear and tear and will have to be replaced. This component of spending is termed depreciation. Net investment subtracts depreciation from gross investment.

Thus, net investment is the increase in the capital stock per period. Net investment could be negative. If firms decide they have too much capital given their expected sales they will invest less than is needed to maintain the size of the current capital stock. Thus depreciation would be larger than gross investment.

What determines the decision by firms on whether to invest? The most elementary explanation is based on the observation that business firms require stocks of capital on hand to produce output. In this context, they must consider two broad aspects of the production decision:

1. Will the size and composition of the capital stock in place allow the firm to produce at a cost level that will achieve its expected target profit rate?
2. Is the capital stock adequate to produce the expected output in the coming periods?

The analysis that business firms engage in to answer these questions then determines their investment decision.

Investment is designed to bridge the gap between the current and desired capital stock which is based on business analysis of future expected output. Investment will be high when the current capital stock is low relative to expected needs. Conversely, investment will be low when the current capital stock is high relative to expected needs.

Given that capital goods last for many years, the capital stock will typically be large relative to current national income (GDP) and current investment (I). Table 25.2 compares the estimated capital stock and GDP for various countries in 2010. The average capital stock–to-GDP (capital output) ratio for the OECD nations in 2010 was 2.4.

Table 25.2 Capital stock and GDP, various countries, 2010, in local currencies

Country	Capital stock	Real GDP 2010	Capital output ratio
Australia	2,855,502,562,904	1,298,899,000,000	2.2
Austria	683,016,438,587	263,318,242,440	2.6
Belgium	807,852,718,331	348,087,000,000	2.3
Canada	3,226,833,382,833	1,325,037,816,083	2.4
Denmark	4,568,552,398,093	1,547,770,000,000	3.0
Finland	362,653,166,133	159,013,000,000	2.3

Country	Capital stock	Real GDP 2010	Capital output ratio
France	4,230,468,965,067	1,774,518,000,000	2.4
Germany	5,737,703,431,790	2,364,092,420,000	2.4
Greece	448,075,712,790	195,587,624,964	2.3
Hungary	35,266,718,116,574	21,807,370,000,000	1.6
Ireland	253,296,673,200	159,914,943,107	1.6
Italy	3,964,191,046,721	1,422,432,146,849	2.8
Japan	2,050,900,350,854,080	540,409,600,000,000	3.8
Netherlands	1,305,843,961,586	550,919,934,367	2.4
New Zealand	352,514,299,624	140,600,000,000	2.5
Portugal	275,424,755,465	162,192,800,000	1.7
Spain	3,387,205,933,615	1,046,329,000,000	3.2
Sweden	6,427,754,866,460	3,300,858,000,000	1.9
Switzerland	1,440,785,679,206	497,771,589,898	2.9
UK	2,322,823,133,929	1,395,312,000,000	1.7
US	27,187,279,574,163	13,087,975,000,000	2.1

Source: Data from OECD.Stat Economic Outlook, Dec 2011, Issue 2, No. 90 (https://stats.oecd.org/index.aspx?DataSetCode=EO90_INTERNET#).

If business firms maintain these ratios then small changes in expected GDP will lead to quite large changes in the required stock of capital goods. Thus, large fluctuations in investment often occur. This observation is the basis of the accelerator model of investment.

25.2 The Accelerator Model of Investment

The simple accelerator model

The accelerator model of investment is based on the observation that increases in GDP lead to increases in the desired stock of capital, which is the amount of capital that business firms would like to have in place given the current and expected economic conditions. An expected increase in GDP which causes an increase in the desired stock of capital will create a gap between the actual capital stock in place and the desired level.

Investment behaviour is characterised by business firms trying to close this gap. The accelerator terminology is based on the premise that small changes in GDP will drive larger changes in investment demand(spending).

We can express the simple accelerator model in terms of the following two equations:

$$(25.1) \quad K^* = vY^e$$

where K^* denotes the desired capital stock, Y^e is the expected future level of GDP, and the fixed multiple v, is the desired capital-to-output ratio (K/Y). For example, if the capital-to-output ratio was 3 and output was expected to increase by $10 billion, then firms would revise their desired capital stock upwards by $30 billion to ensure that they had enough productive capacity in place to meet the expected increase in demand for goods and services. Then we can write the following:

$$(25.2) \quad I_t = [K^*_t - K^*_{t-1}] + I_R$$

where the subscript t is the current time period and $t - 1$ is the previous period (which could be last month, quarter, year, or whatever depending on the frequency of the data). I_R is the replacement investment(depreciation). Thus, Equation (25.2) says that the current flow of investment spending (I_t) is equal to the change in the desired capital stock plus depreciation. We have assumed that the capital stock in the previous period was correctly aligned to the corresponding level of real GDP, i.e. $K_{t-1} = K^*_{t-1}$.

Combining the two equations by substituting K^* from Equation (25.1) into Equation (25.2), we get the simple accelerator model:

$$(25.3) \quad I_t = [K^*_t - K^*_{t-1}] + I_R = v[Y_t - Y_{t-1}] + I_t = v[\Delta Y_t] + I_R$$

This tells us that the flow of investment demand(spending) will be equal to replacement investment(depreciation) plus the desired capital-to-output ratio (v) multiplied by the change in GDP, assuming that expectations about the level of GDP are realised. Remember v is the accelerator coefficient.

While this model of investment is oversimplified (so that we will make it more complicated later in the chapter), it does provide some essential and useful insights that describe investment behaviour in most economies. These insights are:

- Investment demand will be much more variable than GDP because investment (the increase of the capital stock to maintain the desired ratio of capital to GDP) is a multiple of the change in income (GDP), and the multiple is typically greater than unity.
- In an actual economy, investment starts to decline before an economy goes into recession. In other words, we can use turning points in the flow of investment spending to make predictions about the direction of the business cycle. It is clear that when GDP is rising in the early stages of a recovery, then the change in income will be positive and investment will be high. But as the growth in GDP tapers off, the change in income decreases which means that investment starts to decline even though the level of GDP may still be rising.

TRY IT YOURSELF

A practical example will help you understand how this model works. Assume that the accelerator coefficient (v) is equal to 3 and, for simplicity, replacement investment is a constant \$400 per period, despite the capital stock increasing and then decreasing due to the impact of net investment. Assume the economy has an initial capital stock of \$12,000. Table 25.3 shows the operation of the simple accelerator model of investment spending for eight periods. Income initially grows from \$4,000 to a peak of \$4,140 then declines back to \$4,000.

Table 25.3 The simple accelerator model

Period (t)	Expected GDP level (\$)	Expected change in GDP (\$)	Net investment ($v\Delta Y$)	I_R	Gross investment	K^*_{t-1}	K^*_t ($=vY$)	GAP
(1)	(2)	(3)	(4)	(5)	(6) = (4) + (5)	(7)	(8)	(9) = (8) − (7)
1	4,000	0	0	400	400	12,000	12,000	0
2	4,040	40	120	400	520	12,000	12,120	120
3	4,120	80	240	400	640	12,120	12,360	240
4	4,140	20	60	400	460	12,360	12,420	60
5	4,120	−20	−60	400	340	12,420	12,360	−60
6	4,080	−40	−120	400	280	12,360	12,240	−120
7	4,000	−80	−240	400	160	12,240	12,000	−240
8	4,000	0	0	400	400	12,000	12,000	0

You might like to create a spreadsheet to replicate Table 25.3 and then vary the key inputs (expected aggregate demand, the capital output ratio) to see what happens to investment.

Table 25.4 Time path of investment in simple accelerator model

Period	Growth in GDP (%)	Growth in gross investment (%)
1		
2	1.0	30.0
3	2.0	23.1
4	0.5	−28.1
5	−0.5	−26.1
6	−1.0	−17.6
7	−2.0	−42.9

Table 25.4 shows the time path of investment spending described in Table 25.3. You can observe that fluctuations in GDP growth drive much larger fluctuations in investment spending growth.

Figure 25.1 captures the relationship between the change in GDP and gross investment. The increased variability of investment for a given change in GDP is very stark.

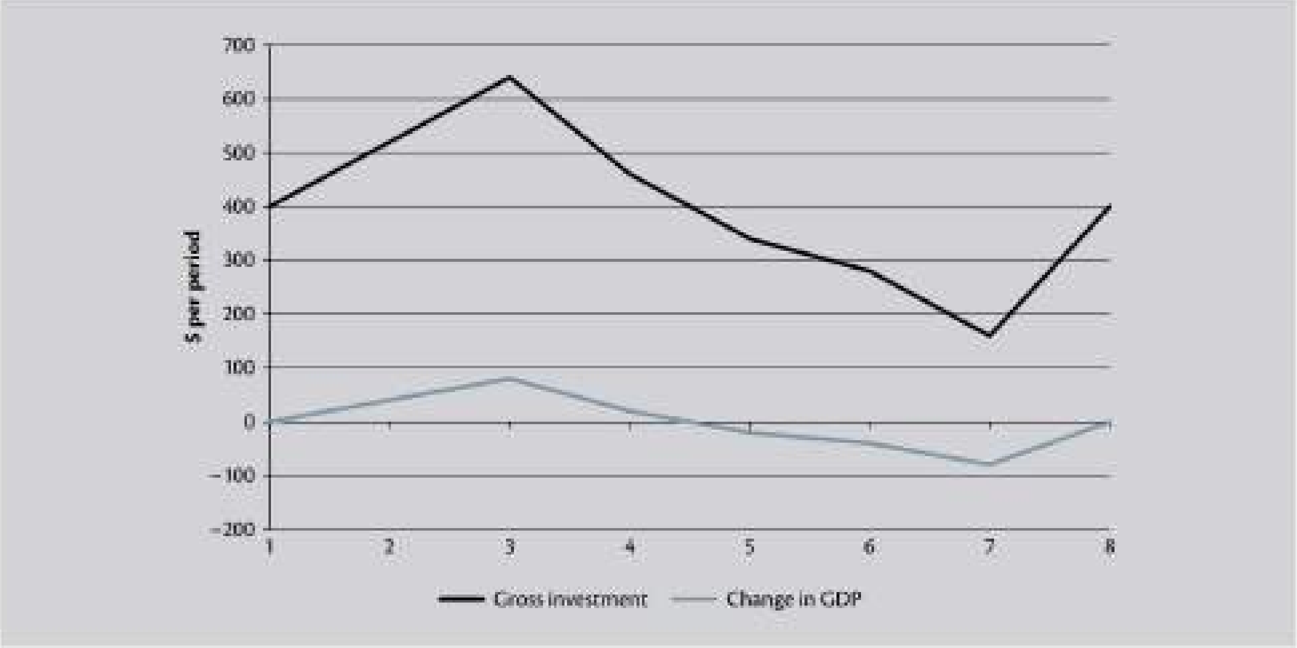
Limitations of the simple accelerator model

While the simple accelerator model is a useful guide it is too simplistic for the following reasons:

- The desired capital stock K^* is unlikely to remain a fixed proportion of GDP. That is, the accelerator coefficient v is likely to be variable in the real world.
- Following a change in GDP, business firms do not attempt to close the gap between their actual and desired capital stocks immediately. That is, like most adjustment processes in the real world, adjustment takes place gradually over time.

These observations led economists to develop the *flexible accelerator model*, which adds some complexity to the explanation of investment.

Figure 25.1 Relationship between change in GDP and gross investment, simple accelerator model



25.3 The Flexible Accelerator Model

The flexible accelerator model of investment behaviour assumes that for a time, business firms can continue to function adequately with a capital stock that is not equal to their desired capital stock.

There are several reasons why firms might want to adjust gradually to their desired capital stock:

- Adjusting the capital stock is very costly especially as suppliers usually charge premiums for faster delivery.
- There are intrinsic time lags involved. It takes time to undertake project evaluation to consider the type of capital, the likely suppliers, the financing arrangements, the delivery details and the installation and training associated with the new equipment.

For these reasons, while net investment is closing the gap between the actual capital stock and the desired capital stock ($K^* - K$), there is incomplete adjustment of the capital stock in the short run.

Business firms employ alternative strategies to maintain production with a less than desired capital stock. These include running extra shifts (that is, employing more labour resources with the existing capital stock by providing opportunities for overtime and extended working weeks).

The flexible accelerator model thus recognises that business firms adjust to increased output by employing a combination of more labour and some investment to augment the capital stock. Over time, they would achieve their desired K^*, although in a growing economy, they are always likely to fall short, given the incomplete adjustment process outlined here.

Rate of adjustment in the flexible accelerator model

The additional question that then arises is what is the speed of adjustment (which we might denote as d)?

The rate of adjustment (d), where $0 < d \leq 1$, is the fraction or proportion of the gap ($K^* - K$) that is closed by the flow of investment per period. If the gap is largely closed (that is, firms get close to their desired capital stock quickly) then d will be close to unity. If the speed of adjustment is slow, the gap persists over a long period.

What does the rate of adjustment depend upon? There are several determinants that economists have identified:

- Adjustment costs: the costs of financing (for example, determined by the interest rate).
- Time factors: how long it takes for new capital equipment to be evaluated, designed, ordered, produced, delivered and installed.
- Economic conditions: interest rates, expected returns on production, animal spirits (business sentiment).

We might say that the higher the interest rates, the higher the costs of financing capital equipment purchases (other things being equal) and so the slower the adjustment will be (see below).

Further, bouts of pessimism will slow down adjustment. We will return to this issue later in the chapter when we consider the asymmetric nature of investment demand.

Implications of incomplete adjustment

By comparison with the simple accelerator model, if in each period the adjustment to the desired capital stock is less than 100 per cent, then the gap between K^* and K will remain non-zero. That is, the actual capital stock will not equal the desired capital stock.

In the flexible accelerator model, the flow of investment demand (spending) is dependent not only on the gap between K^* and K but also the speed of adjustment, d. The following equation captures this formally:

$$(25.4) \quad I_t = d[vY_t - K_{t-1}] + I_R$$

So with $K^* = vY_t$ then the extent to which the gap between the desired and actual capital stock is closed determines the investment flow in addition to the replacement investment (depreciation).

The larger is the coefficient *d*, the closer is the adjustment path of investment associated with the flexible accelerator to that described by the simple accelerator model.

The simple accelerator model posited what we might term an explosion/contraction pattern of investment demand which is clearly unrealistic. The flexible accelerator model predicts a smoother investment path since the adjustment is partial, but still quite variable. Net investment can be sustained even when output growth falls.

Consider the case of a growing economy. Investment spending will be positive because the actual capital stock is below the desired capital stock (K^*). Business firms use investment to adjust towards their desired capital stock.

Table 25.5 compares the net investment adjustment paths for the simple and flexible accelerator using the data of the previous example (see Table 25.3) with an adjustment parameter, $d = 0.3$, which means that firms close 30 per cent of the gap between the desired and actual capital stock each period. It is obvious that the flexible accelerator model exhibits much greater stability.

25.4 Expectations and Interest Rate Impacts on Investment Demand

We saw in the discussion of the flexible accelerator model of investment demand that economic conditions have an impact on the estimates of desired capital stock by business firms. Two factors were identified as being important:

- Expectations of future economic conditions; and
- The interest rate.

Business firms are continually forming expectations about future output. Firms must make resource commitments (working capital, labour and so on) well in advance of realisation (sales) and so the scale of production at any point in time reflects the guesses they make in a highly uncertain world.

Managers wonder whether a change in output that they observe in the current period will be sustained or not. They consider whether observed changes in output are the result of transitory (ephemeral) factors or are likely to be enduring. They wonder whether a rise in demand (output) today may be followed by a fall tomorrow.

The flexible accelerator model of investment is called a partial adjustment model (PAM) and the lack of complete adjustment to the desired capital stock reflects, in part, the uncertainty about the duration of the change in income.

When demand and output are unusually high, business firms might form the view that demand will taper off and so they will allow the current capital stock to depreciate more than usual (that is, net investment will be negative).

Table 25.5 Time path of net investment in simple accelerator and flexible accelerator models, $d = 0.3$; $v = 3$, and replacement investment = 400

Period	Change in GDP	Net investment	
		Simple accelerator	Flexible accelerator
	($)	($)	($)
1	0	0	0
2	40	120	36
3	80	240	97
4	20	60	86
5	−20	−60	42
6	−40	−120	−6
7	−80	−240	−77
8	0	0	−54

Alternatively, when they consider demand and output levels to be unusually low, a reasonable expectation is that demand will rise and so firms may overinvest in the short run to ensure they have enough capital in place to meet their expected future demand. In other words, they create productive capacity that is beyond their immediate requirements.

The role of expectations also helps us understand the asymmetry observed in investment spending across the economic cycle. Investment in new capital stock usually requires firms to make large irreversible capital outlays. Capital is not a piece of putty that can be remoulded at will into whatever configuration might be appropriate (that is, different types of machines and equipment). Once the firm has made a large-scale investment in a new technology, it will be stuck with it for some time.

In an environment of endemic uncertainty, firms become cautious in times of pessimism and employ broad safety margins when deciding how much investment expenditure to undertake. Accordingly, they form expectations of future profitability by considering the current capacity utilisation rate against their normal usage. They will only invest when capacity utilisation exceeds its normal level. As a result, investment spending varies with capacity utilisation within bounds and therefore productive capacity grows at a rate which is bound from below and above. The asymmetric investment behaviour we observe in most nations thus generates asymmetries in capacity growth because productive capacity only grows when there is a shortage of capacity.

This insight has major implications for the way in which economies recover and the necessity for strong fiscal support when a deep recession is encountered. We considered these issues in Chapter 21 *Fiscal Policy in Sovereign Nations*.

What about interest rates? As we learned in Chapters 13 to 15, investment decisions taken by entrepreneurs form an important component of total aggregate demand and income and as a result help explain changes in total output and employment.

Firms are continually making guesses about the future – what the overall state of demand for their products will be, what they are likely to receive by way of revenue if their sales match these expectations, and what it will cost them to produce the output necessary to meet this demand.

Firms also have various choices about what products to produce and how they can produce them (that is, the choice of technique). Firms are driven by the desire to make profit and will thus make choices among different types of productive equipment, on the basis of which will contribute the most profit subject to a range of other considerations, many of which are subjective. For example, a firm that wishes to keep its good standing in the wider community will probably eschew the use of equipment that is damaging to the local environment, even if its use was legal and would generate more profits than other options.

Whether firms use retained profits to fund future investment or seek funds from the markets, there is a cost involved in purchasing new capital. A firm may have retained earnings to invest. It has the choice of investing them in new plant and equipment, or perhaps purchasing financial assets which yield a positive rate of return (for example, a bond). The firm will usually be driven by the need to stay in its present business and therefore defend its market share, which means it will want to use the available funds to purchase best practice, productive infrastructure. However, it may, sometimes, postpone the upgrading of its productive capital.

Investment decisions will thus depend on whether the productive asset being purchased delivers a positive return above the cost. We can build on our understanding of **present value** (PV) to advance this idea in relation to the investment in productive capital, but we need to develop some more technical skills.

25.5 Introduction to Cash Flow Discounting and Present Value

The investment and production decisions of firms are influenced by the expectations that entrepreneurs form about future revenue and cost streams. This enables them to make guesses about future profits.

Different investment options will typically have different time profiles of revenues and costs. For example, one piece of equipment might involve higher outlays in the immediate future but deliver higher returns in a later period compared to another piece of equipment that delivers the same flow of production services, is cheaper in the near future but entails higher costs in the future.

In this section, we cover some introductory concepts that allow us to compare future flows of costs and revenues. The aim is to come up with an approach that can express all future monetary flows in what we call a **present value**, and which provides a means of comparison between investment projects characterised by complex patterns of revenue and costs flows into the future.

Start with a rate of interest (the cost of obtaining funds) of ten per cent. If you had \$100 now and loaned it for a year then at the end of year one you would have \$100 plus the interest earned which is given by the following formula:

$$(25.5)\quad \$100 + 0.10 \times \$100 = \$100(1 + 0.10) = \$110$$

This arithmetic generalises to the following model of simple interest:

$$(25.6)\quad P_{t+1} = P_t(1 + i)$$

where P_{t+1} is the amount received in period $t + 1$ if P_t is the amount you loan at time t, and i is the nominal rate of interest.

We now extend the model to incorporate **compound interest**. What if you were to reinvest P_{t+1} for the second year? You would expect to receive an amount P_{t+2} according to the following formula:

$$(25.7)\quad P_{t+2} = P_{t+1}(1 + i)$$

From Equation (25.6) we know that $P_{t+1} = P_t(1 + i)$ so Equation (25.7) can be rewritten as:

$$(25.8)\quad P_{t+2} = P_t(1 + i)(1 + i) = P_t(1 + i)^2$$

We can generalise this equation to a loan with a duration of n years:

$$(25.9)\quad P_{t+n} = P_t(1 + i)^n$$

Equation (25.9) is a compound interest formula and assumes that the interest is added (compounded) at the end of each year. The formula gets more complicated if there are multiple compounding periods within the year. We do not consider these complications in this textbook.

You will appreciate from the concept of compound interest that a sum invested at a positive interest rate now will grow to a larger future nominal amount in the future. We can deploy this concept in reverse to calculate what the sum of cash that you expect to receive at some future date is worth today. You can see that this type of information is essential for a firm making investment decisions when the cash returns (or outlays) are received (or incurred) at some future date.

In general terms, how much is some cash, P_{t+1}, that you are to receive next year, worth today? From Equation (25.6) we can see that:

$$(25.10)\quad P_t = P_{t+1}/(1 + i)$$

Consider our initial example where P_{t+1} was \$110 and i was ten per cent. Equation (25.10) tells us that \$110 received at the end of the year would be worth \$100 now if the interest rate over the year was ten per cent.

You can think about this in the following way. If you needed cash now and knew you were going to receive \$110 at the end of this year and the current interest rate was ten per cent then how much would you be able to sell your future claim on income for in the open market? The answer is \$100 as long as the receipt at the end of the year was risk free.

No one would be prepared to give you say \$105 now in return for \$110 at the end of the year if the interest rate was ten per cent because they would be losing money. On the other hand, you would be flooded with offers below $100 now if you were selling an asset (income flow) that would deliver the buyer $110 and interest rates were ten per cent.

We also know that the current value of a future cash flow will be sensitive to the rate of interest. Imagine that the rate of interest was to fall during the year to five per cent. Then the future flow of cash of $110 would be worth $104.76, that is *(110/(1 + 0.05))*. You can verify that if you were to loan $104.76 now at five per cent then at the end of year one you would have $110. The future cash flow is less valuable now if the interest rate was to rise.

The current value of a future cash flow is called its *present value* (PV).

Equation (25.10) can be generalised in the same way as Equation (25.9) was generated from the specific case in Equation (25.5). The general formula for the present value (at time *t*) of a cash flow to be received at the end of period *t* + *n* is:

$$(25.11) \quad PV_t = P_{t+n}/(1+i)^n$$

What about the situation where the expected cash flows are distributed across several different time periods? In that case, the present value of a flow at time t, P_t can be written as:

$$(25.12) \quad PV_t = P_{t+1}/(1+i) + P_{t+2}/(1+i)^2 + P_{t+3}/(1+i)^3 + \ldots + P_{t+n}/(1+i)^n$$

where ... refers to terms we have not written between time period *t* + 3 and *t* + *n*.

The conversion of disparate future cash flows across time into present values is generally called **discounted cash flow** analysis.

25.6 Keynes and the Marginal Efficiency of Investment

In Chapter 11 of *The General Theory of Employment Interest and Money*, John Maynard Keynes developed his theory of investment based in his concept of the **marginal efficiency of capital**.

Keynes defined the marginal efficiency of capital as:

> *being equal to that rate of discount which would make the present value of the series of annuities given by the returns expected from the capital asset during its life just equal to its supply price. This gives us the marginal efficiencies of particular types of capital assets. The greatest of these marginal efficiencies can then be regarded as the marginal efficiency of capital in general. (Keynes, 1936: 135–6)*

His definition is difficult to understand and created some controversy, which we will touch upon here.

It is clear that Keynes was thinking about two aspects of the investment decision:

(a) The "prospective yield of the investment" which is a "series of annuities Q_1, Q_2, ... Q_n". These are future flows of cash associated with an investment which the entrepreneur "expects to obtain from selling its output". You can relate these flows to the earlier discussion of compound interest.

(b) The supply price of the asset which is the "price which would just induce a manufacturer newly to produce an additional unit of such assets" or in more simple language, "the replacement cost" (Keynes, 1936: 135).

Keynes' 'rate of discount' is also known as the *internal rate of return* of a project. As we saw in the last section, we can calculate a present value for any future cash flow stream. **A *net present value* would be the present value of the revenue to be received minus the present value of the costs of the project.**

Given the current interest rate a positive net present value for a project corresponds to a project that earns a positive rate of return in current dollars whereas a negative present value means that the project would lose money in current dollars.

The internal rate of return (IRR) is the interest rate that would discount future income and cost outlays such that the net present value (NPV) was zero.

Table 25.6 A simple cash flow for an investment project

Year	Cash flow $
0	−10,000
1	2,500
2	3,200
3	3,500
4	3,300
5	2,500

Consider Table 25.6 which provides data for a specific investment project. In the current year, the firm has to spend $10,000 to purchase the equipment. In subsequent years, it receives the cash flows as indicated. We can assume there is no scrap value for the equipment after Year 5. What is the internal rate of return (IRR) for this project?

The present value of the costs is $10,000 because the costs are all incurred in the current year (that is, now). The dollar sum of the cash returns is $15,000 but as we saw in the previous section, the dollar amounts cannot be compared across time periods because of the impact of compounding.

The present value of the flow of revenue is given by the formula in Equation (25.12). Using the data in Table 25.6 we would write this equation as:

$$(25.13)\quad PV = \$2{,}500/(1+i) + \$3{,}200/(1+i)^2 + \$3{,}500/(1+i)^3 + \$3{,}300/(1+i)^4 + \$2{,}500/(1+i)^5$$

The internal rate of return is the discount rate (i) that satisfies the following equation:

$$(25.14)\quad NPV = -\$10{,}000 + \$2{,}500/(1+i) + \$3{,}200/(1+i)^2 + \$3{,}500/(1+i)^3 + \$3{,}300/(1+i)^4 + \$2{,}500/(1+i)^5 = 0$$

The solution for i requires mathematical techniques that are beyond the reach of this textbook. (For interested students we must solve the roots of Equation 25.14.) However, the solution can be found using spreadsheet functions, and gives $i = 0.151$. In other words, the IRR is 15.1 per cent.

Therefore, the project in Table 25.6 will be profitable if the cost of borrowing funds to fund the project (that is, the market rate of interest) is below the IRR.

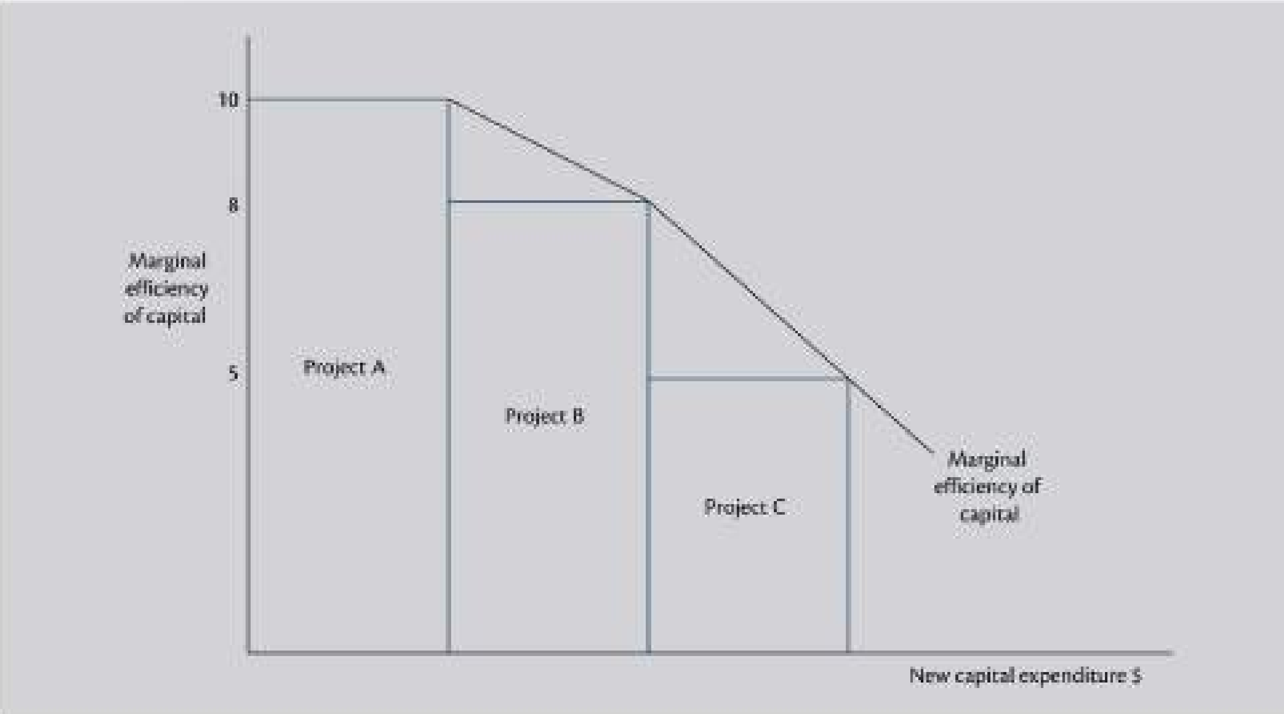
TRY IT YOURSELF

You might like to input the data into a spreadsheet and compute the present value of the revenue stream in Equation (25.13) using a discount rate of 15.1 per cent. You should verify that it is equal to $10,000, which is exactly the present value of the initial outlay. The actual result you get may not equal $10,000 exactly but this is due the approximate iterative solutions used by the spreadsheet.

We can interpret Keynes' concept of the marginal efficiency of capital (MEC) as being the return *at the margin* that a firm would expect to earn by investing in new capital equipment. It is not the market return on existing assets. When considering an investment, MEC is equivalent to the internal rate of return (IRR).

Consider Figure 25.2 which shows three investment projects A, B and C, ranked by their respective MECs. Project A has a MEC of ten per cent, while Project B has an MEC of eight per cent and Project C has a MEC of five per cent.

Figure 25.2 Marginal efficiency of capital and investment projects



The firm must consider how much new capital expenditure it will incur for the coming year. If the market rate of interest is currently nine per cent, then the firm would only be interested in investing in Project A, which means that its capital expenditure in the current planning period would be limited to Project A.

Should the market interest rate drop to below eight per cent, then it will be profitable to borrow sufficient funds (or use retained earnings) and invest in both Project A and Project B. As a consequence, total investment will rise. The firm could expand investment to Project C if the market rate of interest drops below five per cent.

The downward sloping MEC line in Figure 25.2 summarises the investment response of the firm to changes in the market interest rate, although we should appreciate that investment projects involve discrete (lumpy) rather than continuous outlays.

As a result, a simple model of investment emerges whereby total investment in the economy is considered to be a downward sloping function of the market rate of interest as depicted in Equation (25.15):

$$(25.15) \quad I = I_0 - bi$$

where I is total investment, I_0 is some level of investment that is independent of the market rate of interest and b is the sensitivity of investment to the market rate of interest, i.

When considering Equation (25.15) you should always think about what lies behind it in terms of the MEC. The simple investment model, which says that rising market rates of interest lead to lower total investment, is based on an assumption that all other things are equal. But as we have seen, underpinning the concept of the MEC is a comparison between the demand side (expected revenue) and the supply side (the replacement cost).

In a growing economy, it is likely that aggregate demand conditions will improve at times when the market rate of interest rises as the central bank often raises its target rate in an expansion. Increasing aggregate demand will improve the revenue cash flows over time and increase the MEC for each project. In other words, we would not observe investment falling when the market rate of interest rose because the IRR of each project could also be increasing.

Thus, it is important to avoid applying a mechanical interpretation of the concept of the MEC. Keynes, in fact, did not think investment would be very responsive to changes in the market rate of interest, especially when the economy was in recession or boom. Expectations formed by entrepreneurs underpin their MEC calculations. For Keynes it is the expected stream of returns that dominates the investment consideration, not the market interest rate.

When the economy is in recession, entrepreneurs become pessimistic and this would negatively impact on their assessment of the future returns from different projects. Further, with substantial excess productive capacity, firms are unlikely to expand the capital stock even if new investment projects become cheaper as the central bank cuts the market interest rate to stimulate demand.

The extreme optimism that typically accompanies a boom also would reduce the sensitivity of investment to changes in the market rate of interest. With high expected returns, firms would be prepared to pay higher borrowing costs. We could express this enhanced optimism by a shift outwards in the MEC line in Figure 25.2 which would make more projects worth pursuing at a given market rate of interest.

If entrepreneurs became excessively pessimistic then the MEC line would shift inwards and fewer projects would be deemed profitable at a given market rate of interest even if the technical aspects of the equipment were unchanged.

For Keynes then, investment is a very subjective act and responsive to how firms feel about the economy. Even though Equation (25.15) seems simple, the ideas underpinning it are anything but simple!

The concept of the MEC was refined by later economists including Abba Lerner who preferred the term *marginal efficiency of investment (MEI)* (Lerner, 1953).

25.7 Minsky's Model of the Investment Decision

In Chapter 26 *Stabilising the Unstable Economy*, we discuss the work of Hyman Minsky in relation to his theories of financial instability. In this section, we note that Minsky extended Keynes' theory of investment by incorporating finance. This was a key component of Minsky's Financial Instability Hypothesis. He began with a 'two price system' interpretation of Keynes, added the external finance of investment, and included the evolution of expectations and behaviour that would occur over the cycle. In this section, we provide a brief outline of his model.

The two price system

- *Current output prices*: The price system for current output covers the goods and services produced and included in GDP. These are set as 'cost plus mark-up' (see Chapter 16), which ensures that the costs of production are covered in addition to a mark-up to cover overheads, taxes, and interest while leaving net profits for owners of firms. It includes production for consumption, investment, government, and exports. However, for his model of investment, Minsky focused on investment goods, with the current output price set at a level that would induce a supplier to provide new capital assets (plant and equipment). We can call this the 'supply price of capital' similar to the concept of 'replacement cost'.
- *Asset prices*: The price system for assets includes both financial assets, such as stocks and bonds, and real assets, such as plant and equipment. Except for cash, these are expected to generate a stream of income and possibly capital gains. Here Minsky follows Keynes' exposition in Chapter 17 of the *General Theory*. The important point is that for many of these assets, the prospective income stream cannot be known with certainty. Given expected returns, there is a 'demand price' that one is willing to pay to own any particular asset. That price is related to the expected returns but also to the liquidity of the asset, that is, the ease with which a position in the asset can be unwound by selling it for cash.

Note that **investment goods (plant and equipment)** are unusual in that they appear in both price systems: they **are a part of current output, but also are assets that can be held through time to generate income**. This is an important point.

Determination of investment

Figure 25.3 was used by Minsky to show the investment decision. The P_k curve shows the demand price, which is the price one is willing to pay for capital goods based on expectations of returns to ownership. As discussed in the previous section, this comes out of the asset price system. Note that it is horizontal up to a point (I_i), after which it slopes downward for reasons we will discuss in a moment. The P_s curve represents the supply price of capital, which is the price at which producers will supply new capital goods. It, too, is horizontal, to the point I_i.

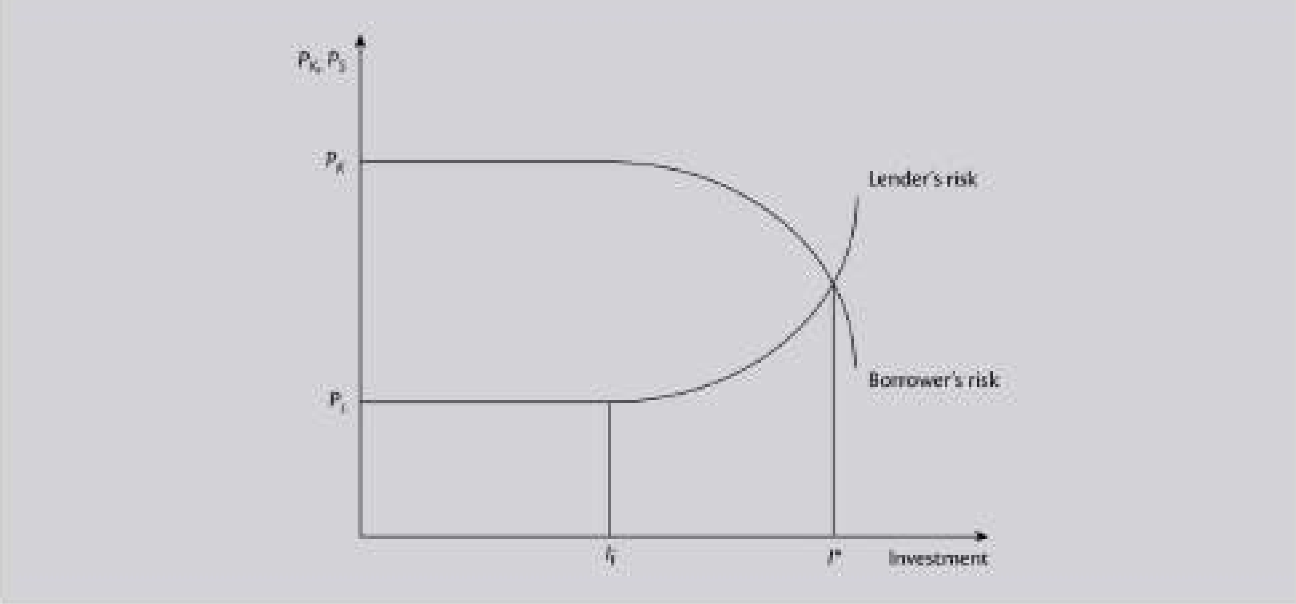
The point I_i is the maximum amount of investment the firm can undertake using its own 'internal' funds, its net revenue from current production. Beyond that point, it must borrow 'external' funds; that is, it must go into debt to finance its position in additional capital goods. From this point, the supply price of capital slopes upward. This is because we must include the costs of finance. More specifically, the finance costs are to cover the lender for 'lender's risk', to compensate for the possibility that the promised repayments are not forthcoming on time. These finance costs include the stated interest rate, but there can be other fees and costs added to the loan terms.

Beyond point I_i, the P_k curve slopes down to represent risk to the borrower of going into debt. If a firm uses its own internal funds to finance capital purchases, it takes a risk that the project will fail to generate expected returns. In a world with uncertainty, the firm can never be sure that any given project will turn out to be a good idea, and the firm's management will certainly regret making bad investments. However, if the firm uses its own funds, it will not be forced into bankruptcy by a bad investment decision. It will lose some funds, and management might be disciplined or even fired, but the firm can survive for another day. When the firm borrows to finance a project that goes bad, it faces the risk that it might default on its debt and be dragged into bankruptcy proceedings. The more it borrows, the greater the 'borrower's risk' that threatens the firm as a going concern.

The intersection of the P_k and P_s curves determines the amount of investment that will be undertaken. Note that if internal funds are used, all we need is for the demand price to be greater than the supply price at the vertical axis. The amount of investment will depend on the lender's and borrower's risks which determine how steeply each curve slopes. If the P_s curve is above the P_k curve, then no investment is undertaken.

Minsky took a dynamic approach in his model of the investment decision. These curves can shift about over the course of a cycle as expectations change. When an expansion gets under way, perceptions of risk might fall and this will flatten both curves if both lenders and borrowers become more optimistic about prospects, so risk assessment declines. Furthermore, the whole P_k curve might shift upward because of greater optimism regarding expected returns from capital investment. Those changes to expectations encourage more investment. However, beyond an economy's cyclical peak, expectations can turn down, when both borrowers and lenders start to worry

Figure 25.3 Minsky's investment decisions



about over-indebtedness, or perhaps about oversaturation of specific markets (that is, too much investment in particular lines of business), or maybe because of a few well-publicised cases of business revenues coming in below expectations. When that happens, both lender's and borrower's risk can rise (which steepens the curves) and the P_k curve can shift downward. Investment then falls. As we will see in the next section, this shift would not be good for profits. Lower investment means lower profits in aggregate (all else being equal). The shift towards more pessimistic expectations becomes self-fulfilling.

25.8 Investment and Profits

What is the relationship between investment spending and the profits that a firm receives? The origin of profits has been an ongoing debate among economists since capitalism succeeded feudal modes of production. In this section, we consider the theory of profits that was developed by Polish economist, Michał Kalecki, one of the early pioneers in developing an understanding of the origin of profits from a macroeconomic perspective.

There are two versions of Kalecki's theory: a simplified version which outlined the fundamental profits equation and the more realistic expanded version which outlined the generalised profits equation.

Kalecki was trained under a Marxist system and so had an advanced understanding of how the production of surplus value predated, and influenced, profit realisation. The mainstream idea is that profits are generated in the 'exchange' process (where goods and services are bought and sold) and reflect the marginal contribution of capital. The Marxist tradition rejects both of these presumptions, arguing that profits are created in the production process by surplus labour, which is appropriated by capitalists. For example, the worker produces enough value to reproduce labour power in four hours, but works an additional four hours to produce surplus value for the capitalist. This is the source of profits; see Chapter 27 *Overview of the History of Economic Thought*.

What Marx didn't show in his approach to profits was how the total volume of profits in a monetary economy was determined in any given period. That was the question that Kalecki sought to answer.

When students first confront the question of the determination of profits from a macroeconomic view they have trouble reconciling it with the accounting concept relating to the measurement of profits at the firm level. As we have learned in earlier chapters, it is easy to fall into the trap of the fallacy of composition when we apply what we observe at the individual level to the macroeconomic level. In this case, attempting to formulate a macroeconomic theory of profits by applying the logic that an individual firm might apply to calculate their profits will lead us astray.

Think about this example. Wage costs are a significant proportion of total costs at the firm level. If an individual firm was to achieve significant reductions in its wage costs, then it might expect to enjoy increased profits. Imagine if all firms simultaneously attempted the same strategy. What would you expect to happen? Given wage costs are also worker incomes and as we have learned, spending is driven by income, we would expect the total revenue for firms to decline as they cut wage costs. There is no reason to expect that overall profits would rise in the economy. Further, investment may also decline as total spending declines, which would further damage the revenue side of the business sector.

The first task then is to determine what factors are important for creating the overall level of profits in the economy.

Kalecki's simplified model

In his simplified model, Kalecki assumed that the economy is comprised of two groups:

- Workers who earn wages and do not save; and
- Capitalists who produce output and earn profits.

He also assumed that the economy is closed and that there is no government sector.

Under these highly simplistic assumptions, Kalecki concluded that 'workers spend what they get' and 'capitalists get what they spend' which means that capitalist profits are determined by capitalists' own propensity to invest and consume, which reverses the way people normally consider the causation. That is, at the aggregate level, **profits are determined by investment** not the other way around.

While Kalecki clearly knew that workers also save, he was able to show that by adopting the restrictive 'workers do not save' assumption the basic insights were not altered but rendered more easy to understand.

To reach that conclusion, Kalecki began with the familiar national accounting identity, which we developed in full in Chapter 4. In his simplified model, the basic aggregate demand equation is written as:

$$(25.16) \quad GNP = C + I$$

where GNP is gross national product or total output and national income, C is total private (household) consumption, and I is total private investment spending per period. These aggregates are all flows of expenditure (and hence income). Private investment is the sum of spending (output) on new productive capital plus changes in inventories.

Note that he used GNP rather than GDP because he assumed the economy is closed. Refer back to Chapter 4 *National Income and Product Accounts* if you need to refresh your memory of the difference between these two aggregates.

In terms of analysing how the total income is distributed, Kalecki assumed that the two 'classes' – workers and capitalists – share the national income so that total wages and salaries (W) plus total profits (Π) equals GNP, so that:

$$(25.17) \quad GNP = W + \Pi$$

where Π is gross profits, which includes depreciation, retained profits, dividends, drawings from unincorporated firms, rent and interest.

Equation (25.16) describes the national product from the spending (demand) side, while Equation (25.17) considers the same aggregate from the perspective of how it is distributed.

If we set the two different views of the national accounts to be equal we get:

$$(25.17) \quad W + \Pi = C + I$$

or:

$$(25.18) \quad \Pi = C + I - W$$

Note that $(C - W)$ is that component of consumption that is attributed to the capitalists (given workers' consumption is equal to W, that is, they 'spend what they get'). Equation (25.19) can then be read as saying that gross profits (Π) is equal to capitalists' consumption $(C - W)$ plus gross investment.

Clearly, the simplified, but fundamental, profits equation was derived from the national accounts and so is true by definition. Kalecki sought to extend his analysis by explaining the causal dynamics that led to the existence of profits overall (at the macroeconomic level), which linked the two sides of the fundamental equation.

The question of interest is which way the causality flows:

- From left to right: Do profits determine capitalist consumption and investment? This is the intuitive way of thinking; or
- From right to left: Do capitalist consumption and investment determine profits?

Kalecki clearly considered the latter causality, that is, that the capitalists 'get what they spend' to be the valid way of understanding profits. He wrote (1965: 45–6):

> *The answer to this question depends on which of these items is directly subject to the decisions of capitalists. Now, it is clear that capitalists may decide to consume and to invest more in a given period than in the preceding one, but they cannot decide to earn more profit. It is, therefore, their investment and consumption decisions which determine profits, and not vice versa.*

He recognised that there is a time lag involved between spending and profits. It is the recognition of this time lag that allowed Kalecki to derive his business cycle model, which we consider later in this chapter.

The essential insight is that investment spending depends on expectations of future aggregate demand that are formed in some prior period. These spending decisions then drive economic activity and so profits are a function of investment in some prior period.

In effect, Kalecki foresaw an accelerator-type process operating. Later in the chapter we will consider how the accelerator interacts with the expenditure multiplier which we derived in Chapter 15, as part of a broader explanation of business cycles.

In the simplified profits model, capitalists' gross saving equals gross investment. Kalecki (like Keynes) demonstrated that the equality of saving and investment is totally independent of the level of interest rates which are determined in the money market. But remember that the simplified model assumes only two sectors: households and firms.

The neoclassical approach presumes interest rates adjust to equilibrate (real) saving and investment and thus ensure that aggregate demand will always be equal to aggregate supply, thus negating the possibility that the economy could suffer from a shortage of demand. This denial of unemployment was the basis of Say's Law (and later Walras' Law).

However, Keynes and Kalecki clearly understood that saving and investment are brought into equilibrium (in a closed economy without a government sector) by variations in national income driven by changes in effective (aggregate) demand. That insight provided the fundamental break with neoclassical thinking that dominated economics (and policy) at the onset of the Great Depression and which, when applied, worsened the depression.

Thus in Kalecki's model, fluctuations in spending drive fluctuations in output and income which ensure the demand drain (saving) comes into equality with spending injections (investment). This is also the basic idea that drives the spending multiplier model. Thus, increased investment spending stimulates aggregate demand and firms respond by increasing production. This, in turn, leads to higher wage and salary payments and higher induced consumption which feeds back into the spending stream and promotes further output and income.

A person looking from a micro perspective might think the profits equation is odd – after all, if the capitalist consumes more, the volume of funds left at the end of some period should be less. It is here that the fallacy of composition enters the fray.

Kalecki asked "what would be the sources of financing this investment if capitalists do not simultaneously reduce their consumption and release some spending power for investment activity?" (Kalecki, 1965: 46). He responded (1965: 46): "It may sound paradoxical, but according to the above, investment is 'financed by itself'."

While that might not be true for an individual business firm, acting in isolation, it can be true for the economy as a whole. This is because the consumption of one capitalist becomes the source of profits for another capitalist. This insight allows us to understand the statement that capitalist investment brings forth its own saving!

Like Keynes (and Minsky), Kalecki understood that 'saving' is not 'finance', the act of saving does not provide the finance needed for investment at the aggregate level. Rather, 'saving' cannot occur until spending has generated the income that can be saved. We must look elsewhere for the source of investment finance.

Kalecki's generalised model

Kalecki subsequently developed a more general model to include a foreign sector, a government sector and a recognition that workers do save. He considered this theory to be applicable to the real world. He examined the influence of the fiscal deficit, the external sector and workers' savings on total profits.

In Chapter 15, we introduced the real expenditure model of national income determination and derived the aggregate demand equation as:

$$(25.20) \quad Y = C + I + G + NX$$

where Y is national income (aggregate spending), G is government spending and NX is net exports (total exports minus total imports).

In Kalecki's model where workers and profit recipients are distinguished, C is taken now to be the aggregate of capitalists' consumption (C_p) and workers' consumption. Workers' consumption is equal to workers' income post tax (W_n) minus workers' saving (S_w). To recognise the different sources of total consumption, Equation (25.20) could thus be written as:

$$(25.21) \quad Y = C_p + (W_n - S_w) + I + G + NX$$

Note that the term $W_n - S_w$ is equal to consumption out of wages and salaries. Total claims on national income (Y) are:

$$(25.22) \quad Y = \Pi_n + W_n + T$$

where Π and W are profits and total wages and salaries, as before, but the subscript n denotes these flows are net of taxes paid, and T is total taxes.

Thus, setting the expenditure components of total income equal to the claims on total income, we get:

$$(25.23) \quad C_p + (W_n - S_w) + I + G + NX = \Pi_n + W_n + T$$

We can solve this for gross profits after tax (Π_n) to get:

$$(25.24) \quad \Pi_n = I + (G - T) + NX + C_p + W_n - S_w - W_n$$

so:

$$(25.25) \quad \Pi_n = I + (G - T) + NX + C_p - S_w$$

which says that gross profits after tax (Π_n) equals gross investment (I), plus the fiscal deficit ($G - T$), plus the export surplus (NX), plus capitalists' consumption (C_p) minus workers' saving (S_w).

Gross profits after tax will be higher, the higher is gross investment (I), the larger the fiscal deficit ($G - T$), the higher is capitalists' consumption (C_p) and the lower is workers' saving (S_w). We do not consider here the behavioural factors that influence C_p and S_w.

Kalecki identified some interesting features of this model. For example, when there are positive net exports and/or fiscal deficits, then gross profits net of taxes (Π_n) will rise to higher levels than would be generated by gross investment and capitalist consumption (as in the simplified model).

Thus, if domestic capitalists are able to increase net exports they will be able to glean extra profits "at the expense of their foreign rivals" (Kalecki, 1965: 51). Kalecki wrote: "It is from this point of view that the fight for foreign markets may be viewed" (1965: 51).

A further elaboration of the foreign sector appeared in Chapter 24 *Policy in an Open Economy*.

Kalecki's generalised model of the determination of aggregate profits showed that fiscal deficits add to capitalist profits through their positive effect on national income. Fiscal deficits lead to the private sector receiving more dollar flows from government spending than it is returning to the government via taxes. Deficits thus provide an increased capacity for capitalists to realise their production plans and sell output because they expand the total aggregate demand in the economy.

Kalecki said that fiscal deficits allow capitalists to make profits (net exports being constant) over and above what their own spending will generate. Government spending not only directly stimulates aggregate demand, but through the multiplier effect it also increases the incomes of household, which in turn, purchase goods and services from firms.

The opposite is the case if the government runs a fiscal surplus, where spending is less than taxation revenue and aggregate profits are reduced. There are two ways in which this occurs. Aggregate spending falls, which reduces the revenue that firms receive. Further, if the surplus is achieved with increased business tax rates, then the firms have less after-tax profit.

A recurrent theme in the public debate, which we will consider in Chapter 30 *Recent Policy Debates*, is the issue of crowding out. We also considered the concept of crowding out in Chapter 21 *Fiscal Policy in Sovereign Nations*.

Many economists think that government spending and private investment compete for a finite pool of saving and this competition must be resolved by higher interest rates, which damages private investment. Accordingly, fiscal deficits are said to 'crowd out' private spending. The same economists typically add a further argument to justify their claim that fiscal deficits are damaging. They allege that public spending is generally wasteful in comparison to private spending because the latter is allegedly 'disciplined' by the market. This is in reference to their belief that the self-regulating market results in the most efficient allocation of resources because inefficient uses of resources are priced out of use by demand and supply forces.

As a preliminary insight into why the crowding-out argument is without substance, we can reflect on Kalecki's profit determination model. The crowding-out argument relies on the claim that savings are finite and borrowers (including firms and government) have to compete with each other to gain access to that finite pool.

Consider Kalecki's conclusion that rising worker saving (lower propensity to consume) and/or falling government deficits impact negatively on profits. The impact is via declines in national income overall. It is probable that when firms are experiencing a reduction in profits as the conditions in the goods and services market deteriorate, they will reduce their rate of investment. On the other hand, equally, private investment adds to private profits and brings forth its own saving via the expansion of national income. In the same way, fiscal deficits add to private profits.

These are fundamental insights of a modern monetary economy that were well understood by Kalecki in his work on the determination of profits and the dynamics of a capitalist economy.

25.9 Business Cycles: Fluctuations in Economic Activity

In the previous section, we considered Kalecki's theory of aggregate profit determination. We gained an understanding of the way in which profits vary as national income fluctuates in response to variations in capitalist consumption and investment, workers' saving, the fiscal balance and the external balance. The fluctuations in economic activity and the resulting changes in national income give rise to the business cycle.

Terminology and patterns

As a matter of terminology, economists often reference an economic variable in terms of the economic cycle. There are three broad relationships:

- **Countercyclical**: which occur when a variable rises(falls) when the level of economic activity falls(rises). That is, we would observe a negative correlation between the magnitude of the variable and the level of economic activity.
- **Procyclical**: which occur when a variable rises(falls) when the level of economic activity rises(falls). That is, we would observe a positive correlation between the magnitude of the variable and the level of economic activity.
- **Acyclical**: there is no relationship between the variable and economic activity. That is, there would be a zero correlation between the two variables.

Typical procyclical variables are household consumption, business investment, imports and employment. Typical countercyclical variables are unemployment and underemployment, and the fiscal balance, which we explained in Chapter 21 when we considered fiscal policy.

Economists also consider macroeconomic variables in terms of the timing of the business cycle. The points in time when the cycle moves from expansion to contraction (the peak) or contraction to expansion (the trough) are referred to as turning points. We consider more complex properties of economic cycles in Chapter 26 *Stabilising the Unstable Economy*.

A variable that demonstrates cyclical behaviour before output has 'turned' is referred to as a *leading* indicator because its movement predates the change in direction of the cycle. Well-known leading indicators include new housing starts, new spending (orders) on plant and equipment by firms, and the purchase of consumer durables by households.

Conversely, a variable that demonstrates cyclical behaviour after output has 'turned' is referred to as a *lagging* indicator because its movement postdates the change in direction of the cycle. Well-known lagging indicators include the rate of inflation, the change in the level of employment, and the rate of wage inflation.

Figure 25.4 depicts a stylised business cycle. The *recovery* or growth phase occurs when real GDP is increasing period after period until it reaches its *peak* and real GDP reaches its localised maximum. The economy then goes into a downturn, sometimes a moderate decline in activity and other times severe, when the real GDP declines overall. If this phase lasts for two or more successive quarters then the economy is said to be in *recession*. At some point the economy reaches the *trough*, which is the lowest point real GDP reaches during that particular cycle. The *trend* real GDP growth rate depicts the underlying direction of real GDP by ignoring the cyclical fluctuations. Note also that the trend real GDP growth rate can rise or fall. In the latter case, it can still remain positive.

Figure 25.5 shows the evolution of annual percentage growth in real GDP for Australia from 1960 to 2015. The formula used to calculate the annualised growth rate from quarterly data is as follows:

$$\textit{Annual growth} = 100 \times (\textit{Real GDP}_t - \textit{Real GDP}_{t-4})/\textit{Real GDP}_{t-4}$$

where the t subscript refers to the quarter in question. So if t was March 2015, then $t - 4$ would be March 2014 and so on.

Figure 25.5 depicts several business cycles of different intensities. The Australian experience is representative of the pattern of economic development of developed economies.

Economic fluctuations are not regular, by which we mean that the time span between peaks and troughs and the depth (amplitude) of the cycle vary over time. While Figure 25.4 depicted peaks that increased over time, it is possible to envisage a peak that falls below the previous peak. The point to understand is that economic activity moves over time in these wave-like patterns, oscillating between peaks and troughs as aggregate demand fluctuates. We refer to a single economic cycle as the time between two peaks because that period contains a complete upswing and downswing.

Other terminology has been used in relation to the economic cycle. For example, economists sometimes differentiate between a recovery (upswing) and a boom in terms of the relationship of real GDP to its trend. So, a recovery (from a trough) becomes a boom once real GDP exceeds its current trend value. Further, a downturn

Figure 25.4 A stylised economic cycle

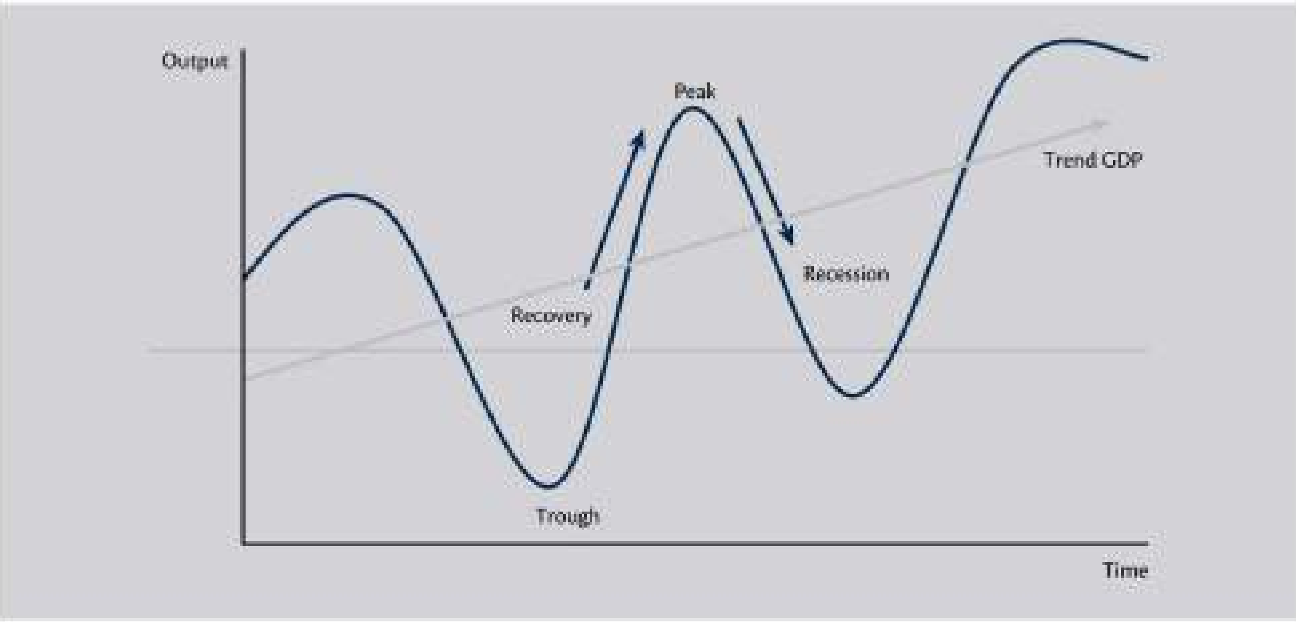
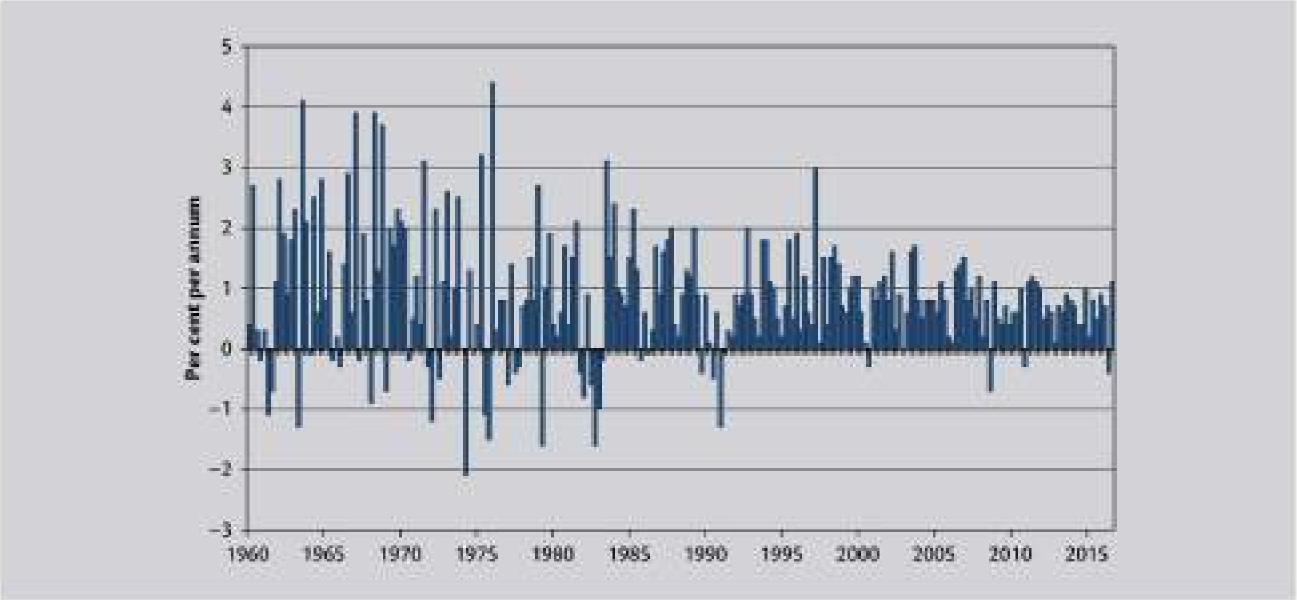


Figure 25.5 Australian real GDP growth, 1960 to 2015, per cent per annum



Source: Authors' own. Data from Australian Bureau of Statistics, National Accounts data.

might describe the fall in real GDP between the peak and the trend line, whereas the economy might be considered to be in recession once real GDP moves below its current trend level.

A very deep and drawn-out recession is sometimes referred to as a *depression* and fortunately, governments now have an array of fiscal policy tools available to them to ensure we rarely encounter recessions so harsh that they can be considered depressions. As the Global Financial Crisis demonstrated though, some governments were reluctant to use these tools because they were ideologically constrained by erroneous free market thinking.

The interaction of the expenditure multiplier and the investment accelerator

In Chapter 16, we introduced the concept of the expenditure multiplier which demonstrated how an injection of spending into the economy would, if there was excess capacity, multiply as the extra income generated was re-spent. In this chapter, we have introduced the accelerator model of investment spending whereby a firm would augment its capital stock through investment spending in order to have enough capital available to produce the expected demand for its output.

In this section, we bring the two concepts together to show how an economic cycle might evolve. The material that follows is considered advanced and is based on the work by American economist Paul Samuelson. The essential idea, however, can be described in a relatively straightforward way.

The accelerator theory of investment spending is based on the notion that net investment is driven by expected changes in output demand. Firms thus seek to install capital equipment embodying the latest technology which will be sufficient to produce goods to meet the expected demand for their output.

The multiplier concept indicates that when there is an exogenous boost in aggregate spending (for example, from government, investment and/or exports) the initial spending increase is multiplied through the expenditure system as consumers are induced by the rising income to increase their consumption.

The two concepts can thus interact. A multiplied spending increase and growth in output will, in turn, increase investment via the accelerator principle. Given that investment is a component of aggregate spending, the rise in net investment will, in turn, have a multiplied impact on total spending and output and so the economy moves into an upward phase in the economic cycle. But once the economy reaches a peak in real GDP, the accelerator becomes a negative influence on net investment which then via the multiplier generates a decline in total spending. **There is thus an interaction between investment as an exogenous driver of the multiplier process and investment as an induced spending reaction as a result of the accelerator principle**.

This was the basic insight that underpinned the Harrod-Domar model of economic cycles (and growth) and supports Keynes' labour market insights whereby an economy has no natural full employment level that it gravitates towards. We considered these insights in Chapter 12, where we learned that this is because the capitalist economy is prone to under-full-employment equilibrium positions.

Recall the simple accelerator model from Equation (25.3):

$$(25.26)\quad I_t = (K^*_t - K^*_{t-1}) + I_R = v(Y_t - Y_{t-1}) + I_R = v\Delta Y_t + I_R$$

This tells us that the flow of investment demand (spending) will be equal to replacement investment (depreciation) plus the desired capital to output ratio (v) multiplied by the change in GDP, assuming that expectations about the level of GDP are realised. Remember that v is the accelerator coefficient. If we only consider net investment we can ignore replacement investment (I_R) because it is likely to change slowly and not be a significant determinant of the business cycle. We also continue to ignore inventory investment to keep the model simple. More elaborate accelerator models consider how firms adjust their inventories to ensure that they can meet unexpected demand changes subject to carrying costs.

To add an element of reality to the model we would hypothesise that when making their net investment decisions, firms respond to changes in output last period ($Y_{t-1} - Y_{t-2}$) rather than those expected in the current period. Further, in the real world, households adjust their consumption with a lag, and so in the following model we use last period's disposable income as the determinant of the household's current consumption.

Combining the acceleration theory of investment with these assumptions, we can write the aggregate expenditure model as:

$$(25.27)\quad Y_t = C_t + I_t + G + NX = cY_{t-1} + v(Y_{t-1} - Y_{t-2}) + G + NX$$

where cY_{t-1} is household consumption, Y_{t-1} is income from the last period, c is the marginal propensity to consume, I_t is net private investment, G is government spending and NX is net exports.

In his original 1939 article, Paul Samuelson simplified this model even further. He adopted a version of the accelerator model in which current investment spending is simply equal to $v(C_t - C_{t-1})$, that is, investment spending is proportional to the change in consumption spending in the current period. Thus, firms ensure that there is sufficient productive capacity to meet the growth in consumption.

He also assumed $c = 0.5$, $v = 1$, $G = 1$ in each period, and $NX = 0$. Accordingly, the output equation is:

$$(25.28)\quad Y_t = C_t + I_t + G_t$$

which, based on the assumed behaviour and parameters, becomes:

$$(25.29)\quad Y_t = cY_{t-1} + v(cY_{t-1} - cY_{t-2}) + G_t$$

since $C_t = cY_{t-1}$.

If we consider period zero, then:

$$(25.30)\quad Y_0 = cY_0 + G_0$$

since output is assumed to be unchanged so that investment is zero.

We can see the changes in total output and its components from period zero as the economy adjusts to an external spending shock by subtracting Equation (25.30) from Equation (25.29):

$$(25.31)\quad Y_t - Y_0 = (cY_{t-1} - cY_0) + v(cY_{t-1} - cY_{t-2}) + 1$$

Table 25.7 Changes in output and its components from a permanently higher level of government spending

Period	Government spending $(G_t - G_0)$	Consumption expenditure $(C_t - C_0)$	Investment expenditure $(I_t - I_0)$	Total GDP $(Y_t - Y_0)$
1	1	0.000	0.000	1.000
2	1	0.500	0.500	2.000
3	1	1.000	0.500	2.500
4	1	1.250	0.250	2.500
5	1	1.250	0.000	2.250
6	1	1.125	−0.125	2.000
7	1	1.000	−0.125	1.875
8	1	0.938	−0.063	1.875
9	1	0.938	0.000	1.938
10	1	0.969	0.031	2.000
11	1	1.000	0.031	2.031
12	1	1.016	0.016	2.031
13	1	1.016	0.000	2.016
14	1	1.008	−0.008	2.000
.....	.....	....	....	....

That time path will be sensitive to the values of the parameters. Samuelson's 1939 example assumed that $c = 0.5$ and $v = 1$, with $G_t - G_0 = 1$ $(t = 1,2,...)$.

Table 25.7 documents the changes in total output and its components as the economy adjusts to a permanent one-unit increase in G, showing the adjustment path in output based on Samuelson's assumptions. The initial introduction of one unit of additional government spending in period one merely increases total GDP by the same amount, given that consumption is assumed to depend on last period's GDP. The accelerator process is also dormant because consumption is unchanged.

The rise in GDP of one unit in period 1, induces 50 cents **extra** consumption in period 2, which triggers the accelerator mechanism such that investment rises by 50 cents, with both changes adding to the persistent higher government spending to generate a level of GDP in period 2 which is two units higher than in period 0.

Without the accelerator mechanism, induced consumption would steadily decline over time after the initial unit increase in G and GDP would come to rest at a higher value determined by the expenditure multiplier. However, with the accelerator present, we observe a peak in GDP after three periods followed by a trough in period 7, and another peak in period 11, and so on.

Thus the introduction of interaction between the multiplier and the accelerator introduces an economic cycle. Different cycles are revealed with different values of c and v. You might like to construct your own spreadsheet to represent this simple interactive model and then vary the values of c and v to see the economic cycles that emerge.

In sum, introducing the accelerator notion to a model with a spending multiplier generates cyclical behaviour. It turns out that choosing some parameter values can lead to fairly stable cyclical behaviour, while different choices lead to explosive cycles. The point is that as we introduce what appear to be reasonable assumptions regarding investment behaviour (that is, making investment a function of consumption), there is no reason to believe that an economy moves toward equilibrium. Rather, the accelerator might lead to cyclical instability even under quite reasonable assumptions. Hyman Minsky took the results from this model and developed a theory of economic instability which we discuss in Chapter 27 *Overview of the History of Economic Thought*.

Conclusion

In Keynes' approach, investment plays a role as a driver of the economy. As we learned in Chapter 16, investment (along with other autonomous spending) has a multiplier impact on aggregate demand. Investment, in turn, is subject to substantial fluctuation because it depends on expectations regarding an uncertain future. In this chapter we found that investment also acts as an accelerator if it is influenced by current flows of spending and income. This insight compounds the effects of investment: if expectations are optimistic(pessimistic), then investment rises(falls), which causes income to rise(fall) through the multiplier effect (as changes to income cause consumption to change) and stimulates more(less) investment, which also has a multiplier effect

In addition, we examined Minsky's approach to investment, which introduces financial considerations into the investment decision. This adds further destabilizing impacts of investment. For example, if investment is proceeding at a rapid pace, thereby boosting economic growth that will tend to produce ever-more optimistic expectations that encourage even more investment. At least some investment is externally financed by borrowing and, as growth picks up, both borrowers and lenders discount the riskiness of debt, which allows investment to proceed at a rapid pace. However, as debt rises, the potential for financial problems increases.

We also introduced the Kaleckian view of the relation between investment and profits. We examined the Kalecki identity that shows that all else being equal, a rise of either investment or the government's spending deficit will create more profits. That, too, can encourage even more investment and produce a euphoric boom. However, should expectations turn around, everything works in the opposite direction: investments fall, profits fall, and firms have trouble making their payments on the debt that has increased over the course of the boom. A financial crisis becomes ever more likely.

The conclusion that we draw based on this analysis is that it is unlikely that growth in a capitalist system would proceed along a stable path. Cyclical behaviour is virtually inevitable and that provides a positive role for government to play in using policy to attenuate the cyclical extremes.

References

International Monetary Fund (IMF) (2016) World Economic Outlook Database, October.

Kalecki, M. (1965) *Theory of Economic Dynamics: An Essay on Cyclical and Long-run Changes in Capitalist Economy*, 2nd edn, London: George Allen & Unwin.

Keynes, J.M. (1936) *The General Theory of Employment, Interest, and Money*, London: Macmillan, 1957 Reprint.

Lerner, A.P. (1953) "On the Marginal Product of Capital and the Marginal Efficiency of Investment", Journal of Political Economy, 61(1), 1–14.

Organisation for Economic Co-operation and Development (OECD) (2011) Economic Outlook, Volume 2011, Issue 2, No. 90.

Samuelson, P.A. (1939) "Interactions Between the Multiplier Analysis and the Principle of Acceleration", *Review of Economic Statistics*, 21(2), 75–8.

Endnote

1. Keep in mind that investment also includes residential housing, which is not considered to be capital, although it is a stock of housing. Most of our focus in this chapter will be on investment in the capital stock used by business to produce output.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

26 STABILISING THE UNSTABLE ECONOMY

Chapter Outline

Learning Objectives

- Develop an understanding of the main heterodox theories of the business cycle.
- Operationalise the insights of Samuelson's multiplier accelerator model that was studied in Chapter 25.
- Understand why neoclassical theories that presume market processes are equilibrium seeking do not apply to real world capitalist economies.
- Be able to criticise the pre-GFC mainstream view of the 'Great Moderation'.
- Begin to understand why it is a mistake for policymakers to rely on the 'invisible hand' to produce economic and financial stability.
- Be able to explain why 'stability is destabilising'.

26.1 Introduction

In the early 2000s mainstream economists and policymakers who followed orthodox macroeconomics congratulated themselves for creating what they perceived to be a new era of economic stability.

Ben Bernanke (who was soon to become Chairman of the US Federal Reserve) even wrote a paper in 2004 referring to the era of the "great moderation". The stability was attributed by the mainstream to a combination of greater reliance on free markets, substantial deregulation of the financial sector, sophisticated financial innovations to reduce risk, globalisation to diversify portfolios, greater fiscal constraints, and greater reliance on central banks that had adopted the New Monetary Consensus approach to policy formation. Just three years later, the world's financial markets collapsed and the Global Financial Crisis (GFC) began.

By contrast, in 1982 Hyman P. Minsky had published a book titled *Can "It" Happen Again?*. By 'it' he meant another financial crash and Great Depression like the one that impacted much of the world in the 1930s. At that time, he concluded that given the institutional constraints put into place by the US's New Deal (and by similar responses in other countries), 'it' was unlikely.

However, over the decades that followed the publication of Minsky's book, the New Deal constraints were slowly removed, freeing the financial system to engage in the same sort of risky practices that had caused the Great Depression. Minsky began to worry that 'it' *might* happen again. While he died in 1996, a decade before the crash in 2007, his analysis presciently identified why 'it' *did* happen again.

In this chapter, we examine the causes of economic instability, with a special focus on *financial* instability. In Chapters 31 and 32, we will address the GFC in more detail and its implications for economic theory and policymaking.

26.2 Economic Cycles and Crises

Economists have long been concerned with the economic fluctuations that occur more or less regularly in what is called the economic cycle (Sherman, 1991; Wolfson, 1994). In Chapter 25, we provided a brief introduction to the concept of an economic cycle. We now outline a typology of the economic cycle and associated theories.

Economists have identified several different kinds of cycles, ranging from the Kitchin cycle (tied to inventory swings and lasting on average 39 months) to the Juglar cycle (lasting seven or eight years and linked to investment in plant and equipment) to the Kuznets cycle of twenty years (associated with demographic changes) and finally to the Kondratieff long wave cycles attributed to major innovations (such as electrification and the automobile) (Kindleberger, 1989). We call the shorter cycles (Kitchin and Juglar) economic cycles, while the longer cycles (Kuznets and Kondratieff) are long wave cycles. Financial factors are likely to play only a small role in some of these longer-run fluctuations.

Generally, economists studying financial instability have tended to focus on periodic financial crises that coincide with the peak of the shorter economic cycle, although financial crises (especially in recent years) can occur at other times during the cycle.

Furthermore, an economy might be financially unstable but manage to avoid a financial crisis. It is best to think of financial instability as a tendency rather than as a specific event, although the typical financial crisis might be the result of unstable financial processes generated over the course of an economic cycle expansion.

In this chapter, we will be concerned primarily with economic instability that has its roots in a financial cause. We are less interested in either economic fluctuations that are largely independent of finance, or in isolated financial crises that do not spill over into the economy as a whole.

A variety of explanations of the causes of financial instability have been offered. One possible cause could be a speculative 'mania' in which a large number of investors develop unrealistic expectations of the profits to be made. They borrow heavily to finance purchases of assets and drive up their prices. Eventually, the mania ends, prices collapse, and bankruptcies follow (Kindleberger, 1989). The tulip mania of 1634, the South Sea bubble in 1719, and the Dotcom boom of the late 1990s can be cited as examples of speculative manias. These booms often develop as, and are fuelled by, fraudulent schemes. Examples of financial crises in which fraud played a large role include the collapse of the Albanian national pension system (1990s), as well as the American Savings and Loan fiasco (1980s) and the US subprime mortgage boom and bust of the 2000s.

Other, mostly mainstream, explanations focus on a sudden interruption of the supply of money or credit that prevents borrowing and forces spending to decline, precipitating a cyclical downturn.

The Monetarist approach attributes financial instability and crises to policy errors by central banks (see Chapter 24). According to this doctrine, when the central bank supplies too many reserves, the money supply expands too quickly and this fuels a spending boom. If the central bank then overreacts to the inflation that this is believed to cause, it reduces the money supply and causes spending to collapse.

Others advance a 'credit crunch' thesis according to which lenders suddenly reduce the supply of loans to borrowers, either because the lenders reach some sort of institutional constraint or because the central bank adopts restrictive monetary policy (as in the Monetarist story).

Finally, one could add exchange rate instability and foreign indebtedness as a precipitating cause of economic instability, especially in developing nations since the breakup of the Bretton Woods system. If a nation owes debt in a foreign currency, when its exchange rate depreciates, this increases the debt burden (since debtors have to provide more of the domestic currency to obtain the foreign currency that is needed to service debt), which causes default and crisis (see three examples of this type of crisis in Chapter 31 *Recent Policy Debates*).

Thorstein Veblen (1904) developed a theory of the economic cycle consistent with his 'theory of business enterprise'. In an expansion, credit is extended on the basis of collateral, which is the capitalised value of assets of the firms, approximated by the market value of the firm's stock. As the stock market booms, a firm's market value is rising, which increases its access to credit.

The market value of a firm includes 'goodwill', an accounting term that applies to 'intangible' assets such as the value of a brand name. As profits grow over the course of expansion, profit expectations grow, increasing the value of the firm, further increasing the ability to get credit. Veblen also argued that the firm's management can actively manipulate stock prices through a wide variety of procedures, which today would include corporate stock buy backs. However, as expansion proceeds, costs begin to rise.

Generally, aggregate demand does not keep up, largely because worker income is constrained, which limits consumption growth. Profits do not meet expectations, causing the market value of firms to fall.

Some firms might have to liquidate assets to meet debt commitments, access to credit becomes more difficult, and stock prices collapse. A crisis ensues. The response by businesses can make it worse: slashing employment and production while trying to maintain prices is self-defeating for the macro economy since this depresses demand and thus makes it difficult to sell output at the desired prices.

Wesley C. Mitchell, a follower of Veblen, has been given credit for making major contributions to studies of the economic cycle. He was one of the founders of the NBER in 1920 and served as its Research Director until 1945. Based on his empirical findings, he outlined the typical economic cycle.

As the economy begins to recover from a depression, profits rise because costs only slowly begin to rise. Beyond some point, however, costs rise faster. Also, long-term interest rates begin to rise, which induces firms to shift to shorter-term loans. Higher costs and interest rates squeeze profits. Firms retrench their spending to maintain solvency; borrowing also falls except by firms that cannot otherwise meet their payments.

Troubled firms are also forced to sell assets. This can degenerate into a financial panic as forced sales cause asset values to plummet. With the crisis and downturn, defaults reduce indebtedness. Wages and other costs decline in the depressed conditions. Marginal firms are eliminated. Inventories are sold off and with little investment, depreciated capital is not replaced.

Eventually the stage is set for a recovery and the economic cycle begins anew. While Mitchell conceded that every economic cycle has its own idiosyncratic features, he argued that in general terms they follow this outline. While he was not known for formulating a *theory* of the economic cycle, many economists followed his lead in using empirical evidence in support of their theories, by often featuring 'stylised facts' in their explanations of the cycle.

The following sections consider alternative heterodox theories of crises, beginning with the Marxist perspective. In Marx's view, crises are inevitable due to internal contradictions inherent in the capitalist system. Indeed, he believed that these crises would become increasingly severe until workers would rise up in revolution to overthrow capitalism and replace it with socialism.

We next turn to the Keynesian theory of crisis (note that we have also covered Keynes' ideas in detail in Chapters 12 and 13). While less critical of capitalism than Marx, Keynes also saw the tendency towards crisis as inherent to capitalism. His followers pay particular attention to Chapters 12, 17 and 24 in the *General Theory* which focus on the *fundamental faults* of capitalism – the tendency to operate below full employment and to generate excessive inequality – which in turn are related to fundamental uncertainty and the preference for liquidity.

The penultimate section will look in detail at the work of Hyman P. Minsky, who extended Keynes' approach to develop the *financial instability hypothesis*, which garnered a great deal of attention after the GFC. The final section will provide a summary of the differences between heterodox and orthodox approaches to cycles and crises.

26.3 Marxist Theory of Crisis

Some analyses have identified processes inherent to the operation of capitalist economies to explain the incidence of crisis. In other words, rather than looking to fundamentally irrational manias or to 'exogenous shocks' emanating from monetary authorities, these approaches attribute crisis to internal or endogenous factors.

Karl Marx in *Capital* claimed that the "anarchy of production" is an inevitable characteristic of an unplanned economy in which decisions are made by numerous individuals in pursuit of profit. The capitalist system is such a system and is subject to "disproportionalities" of production, so that some of the produced goods cannot be sold at a price high enough to realise expected profits.

Central to his explanation was the recognition that production always begins with money, some of which is borrowed. This is used to purchase labour and the instruments of production in order to produce commodities for sale. If however, some of the commodities cannot be sold at a sufficiently high price, loans cannot be repaid and bankruptcies occur.

Creditors may also be forced into bankruptcy when their debtors default because the creditors themselves will have outstanding debts they cannot service. In this way, a snowball of defaults spreads throughout the economy, generating a panic as holders of financial assets begin to worry about the soundness of their investments.

Rather than waiting for debtors to default, holders of financial assets attempt to 'liquidate' (sell) assets to obtain cash and other safer assets. This high demand for 'liquidity' (cash and marketable assets expected to hold their nominal value) causes prices of less-liquid assets to collapse, and at the same time generates a reluctance to spend as people try to hoard money. Thus, a financial crisis occurs in conjunction with a collapse of aggregate demand (Sherman, 1991; Marx, 1990; 1991; 1992).

Marx argued that money is a precondition for an economic cycle because it allows the separation of sale and purchase. He criticised Ricardo for denying the possibility of a glut. Ricardo's mistake was viewing capitalist production as similar to production for barter exchange. No one would offer something for barter exchange without simultaneously demanding something in return: 'supply creates demand'. But if one can sell for money, then a sale does not necessarily create a demand. Ricardo's theory cannot apply to a capitalist economy, where production is for sale for money.

Marx argued that even in the case of 'simple commodity production' (C-M-C') crisis becomes possible: one produces a commodity for sale for money (C-M) in order to purchase another commodity but one might choose to postpone the next purchase (M-C') by holding on to that money instead. Here C stands for a commodity that is produced, M is the money for which the produced commodity is sold, and C' is the (different) commodity that the first producer purchases with the money from the sale of the first commodity (C). In this sequence, money is simply an intermediary, a medium of exchange, as the purpose of production is to obtain commodities (C') in exchange. One might hold money temporarily as a store of value until one finds the desired commodity (C'). This is still not a theory of capitalism, but even simple commodity production can experience a crisis.

As we saw in Chapters 3 and 11, in capitalism, production begins with money (M) only on expectation of receiving more money (M') later. In this case, many things can go wrong. Producers may not believe production will be profitable so it is not undertaken at all. Alternatively, produced goods may not be sold because capitalists have misperceived the demand for the commodities they produce, or even if products are sold, the profits may be lower than expected.

Marx defines capitalist production as the process of M-C-P-C'-M' (where P is the production process that transforms commodity inputs (C) to final commodities (C') for sale to realise more money (M') which includes profits. Note here that C is used to indicate commodity inputs and C' is used to denote final commodities. The capitalist producer does not have a choice to *not* sell. Capitalists must sell in order to recoup the M with which they started. They do not produce commodities for their own use, but rather produce them only to sell at profit.

Note also that Marx was not interested in a 'swindler' theory of profit, that is, a theory based on a 'buy low, sell high' strategy. Prior to Marx, this was a common explanation of the source of profits, that is, the argument that profits are created out of trade. However, Marx wanted to explain the source of profits in the production process itself. In other words, his question was: *Will all capitalists be able to achieve M'?* He concluded that this is possible only if aggregate demand is sufficiently high and properly distributed. If not, any number of things can go wrong to precipitate a crisis, including overproduction of commodities or a sudden fall of the profit rate. Those comprise 'actuality theory' – what actually happens. But all we need is a theory of effective demand based on a monetary theory of production to allow for the possibility of crisis.

Marx argued that capitalism is generally subject to increasingly severe crises. Briefly, this is due to the tendency to replace 'living labour' (human labour) with 'dead labour' (capital equipment produced by labour). In Marx's theory, only living labour can produce surplus labour value ('dead labour' can only reproduce the labour required to produce the capital equipment), so as dead labour replaces living labour, the amount of surplus value produced tends to fall. Surplus labour value is the source of profits and as surplus value falls, profits also fall. While Marx enumerated countervailing tendencies that could ameliorate and postpone this, he believed that the rate of profit will tend to fall over time. As profit falls, capitalists have less reason to produce, which generates downturns brought about by insufficient demand. The downturn makes it difficult to realise profits and service debts, which can generate an increase in demand for credit to be used to pay debts. A fire sale of assets can accompany this demand for credit, fuelling falling asset prices. The downturn can turn into a severe financial crisis. Marx believed that capitalism is ultimately doomed, as (eventually) workers would rise up in response to these increasingly severe crises, overthrowing capitalism and replacing it with socialism.

Baran and Sweezy (1966) replaced Marx's concept of a tendency of the falling rate of profit by a tendency of the surplus to rise. This occurs as capitalists squeeze workers, increasing exploitation (reducing wages or requiring more work hours for the same pay). The surplus must be absorbed through: (a) increased consumption; (b) investment; and (c) waste. But as capitalist profit rises, their consumption as a ratio to profit falls; and the aggregate supply effect of investment generates excess capacity so investment falls. Thus, waste is the only solution, largely taking the form of sales effort, military adventures, and outsized growth of the financial sector to absorb the excess surplus created.

In a similar vein, Joseph Gillman (1957) argued that modern capitalism produces an uninvestable surplus. In a sense, the problem is that capitalism is too successful: its capacity to produce exceeds its ability to find new areas in which to invest. Capital-saving technological advance reduces the need for surplus, while rising productivity of labour increases the surplus. Some of the surplus can be absorbed in wasteful spending (some private and some public) such as advertising and war. However, this is limited because it creates social, moral, and political tendencies that are incompatible with democracy as the wasteful spending drains the public purse. Hence, while a policy to increase demand can work temporarily to create uses for the uninvestable surplus, it cannot work in the long run. It requires an ever-growing public sector to absorb the surplus, but capitalists will not sit back and let the economy become socialised.

Along similar lines, Kalecki (1954) argued that mature capitalism exhibits a tendency to produce excess capacity that then leads to stagnation. The demand for consumption goods is constrained by total wages (paid in both the investment and the consumption goods sectors). The demand for investment goods is a function of realised profits. Rapid accumulation of capital requires that the demand for capital grows faster than total wages, but excess productive capacity limits the demand for new capital equipment. Hence, capitalism needs another source of 'spontaneous' demand, such as technological innovations to absorb capital, or the rising growth of government spending. The problem is worsened by a steady rise of the degree of monopoly power since the capacity to set prices is supported by lower output and the maintenance of excess capacity (to ward off competition). That depresses demand and only worsens the problem as the economy stagnates.

Steindl (1952) argued that the tendency toward stagnation implies that the rates of growth of capital and of profit fall, reducing the share of capitalists. Monopoly capital tries to maintain prices in the face of low demand by cutting production, leading to even more excess capacity; stagnation results in order to maintain a high profit margin, but is self-defeating because unemployment and stagnant wages limit sales.

In a famous paper, Kalecki (1943) argued that 'Keynesian' aggregate demand policies to address rising unemployment caused by stagnation would produce a capitalist reaction, given that capitalists see unemployment as a means of suppressing wage demands from workers. However, Kalecki did not foresee the political power of social democratic governments that allowed them to pursue full employment policies in the post-Second World War period, despite ongoing resistance from capitalists.

What is important in all of the approaches inspired by Marx is that the tendency to crisis is inherent to the dynamics of the capitalist system. Crises are not due to irrational exuberance, policy mistakes, or uncertainty.

26.4 Keynesian and Post-Keynesian Theories of Crisis

Some of the same conclusions from Marx's analysis were also reached by Irving Fisher (1933) in his "debt deflation" theory of the Great Depression, as well as by John Maynard Keynes (1936) in *The General Theory*. While Fisher devised a theory of special conditions in which markets would not be equilibrating, in Keynes' theory these are general conditions operating in monetary economies.

Briefly, Fisher attributed the severity of the Great Depression to the collapse of asset prices and the ensuing financial crisis that resulted from an avalanche of defaults (Fisher, 1933; see also Galbraith, 1972). As output and income fell, this made it hard to service debts, forcing asset sales to try to raise the funds needed to make payments. However, with everyone trying to sell assets in 'fire sales', prices plummeted. At the low prices, even more assets would have to be sold to cover debts. Debtors were 'under water' – the nominal value of their debt exceeded the value of their assets – leading to bankruptcy. It was this combination of 'debt deflation' (falling asset values) plus the severe economic downturn that made the Great Depression so terrible.

Adopting an approach similar to Marx's *M-C-M'* notion that capitalist production of commodities begins with money (*M*) on the expectation of ending with more money (*M'*) later, Keynes developed a general theory of the determination of equilibrium output and employment that explicitly incorporated expectations (see Chapter 13). He concluded that there are no automatic, self-righting forces operating in capitalist economies that would move them toward full employment of resources.

Indeed, he described destabilising 'whirlwinds' of optimism and pessimism, in striking contrast to the Smithian notion of an 'invisible hand' that would guide markets toward stable equilibrium. Also, like Marx, Keynes identified what he called the 'fetish' for liquidity (or money) as a primary destabilising force that erects barriers to the achievement of full employment. In summary, rising liquidity preference lowers the demand for capital assets, which leads to lower production of investment goods and thus falling income and employment through the multiplier effect. In Chapter 12 of the *General Theory*, Keynes attributed the swing between boom and bust to the effect of these whirlwinds on investment.

REMINDER BOX

Recall from Chapters 12 and 13 of this textbook, that in Keynes' approach, investment is a function of the prospective yields to be generated by capital equipment. These yields are uncertain as they will be generated in the future. The investor weighs up the expected yields (formulated as the marginal efficiency of capital) against the interest rate. If the marginal efficiency of capital is greater than the interest rate, investment is undertaken. That then has a multiplier impact on output and employment.

In the orthodox interpretation of Keynes, it is the fluctuation of the interest rate that causes investment, and hence the economy as a whole, to exhibit fluctuations. This explains in part why monetary policy is argued to be the main stabilisation tool for the macroeconomy. Keynes, however, argued that it is fluctuation of the marginal efficiency of capital that drives the cycle, with estimates of the marginal efficiency of capital being high during periods of optimism and low, even negative, when pessimism reigns.

Why is the marginal efficiency of capital susceptible to such large and destabilising movements? Because we cannot know the future, and we know that any expectations we hold are weakly grounded because of our uncertainty about the prospects for investment. In Chapter 12 of the *General Theory*, Keynes distinguishes between two types of expectations that go into determining the prospective yield (1936: 147–8):

(a) Expectations that depend on *existing* facts that can be more or less known; these depend in turn on the existing stock of types of assets, and the strength of current consumer demand. These are *short-term* expectations. The entrepreneur estimates the price that they will get for a product, using the existing plant and equipment, if production begins today.

(b) Expectations that depend on *future* events that cannot be known but can only be forecast with more or less confidence. These are *long-term* expectations. They depend on the 'most probable forecast', with a discount according to the confidence we hold in that forecast. Long-term expectations are subject to the uncertainty that we have about future changes in the type and quantity of the stock of assets, changes in consumer tastes, strength of effective demand throughout the life of the assets, and changes in the wage unit.

The short-term expectations apply to the use of existing plant and equipment, while long-term expectations govern decisions to invest in new plant and equipment. The investment decision, by its very nature, depends on loosely held expectations that are subject to change:

> *The outstanding fact is the extreme precariousness of the basis of knowledge on which our estimates of prospective yield have to be made. Our knowledge of the factors which will govern the yield of an investment some years hence is usually very slight and often negligible. (Keynes, 1936: 149)*

How can investors cope with such a high degree of uncertainty? In normal times they fall back on convention, presuming that current conditions will continue into the future; in other words, they rely on 'existing facts' and extrapolate those into the future. If everyone behaves this way, even though they recognise there is no basis for such a belief, it will help to stabilise investment behaviour which will tend to produce economic stability. The problem is that because we know that our knowledge of the future is precarious, it does not take much adverse news, say poor macroeconomic data, to dislodge the convention.

Keynes argued that the development of the stock market helped to create conditions that made it more likely that the investment decision would become infected by whirlwinds of optimism and pessimism. With the rise of publicly traded corporations, ownership and management of firms became increasingly separated. Ownership was widely spread among thousands of holders of stocks, most of whom knew very little about the business of the firm and hence relied on bits of news and even wild rumours when deciding whether to buy or sell stocks.

Keynes (1936: 154) referred to the tendency of the stock prices of firms manufacturing ice to rise in the summer and fall in the winter as an example of the "excessive, and even an absurd, influence" of short-term profitability on share prices. Today one could just as well refer to the excessive influence on the stock market of rumours about a dotcom firm, or a news report about some new drug developed by a big pharmaceutical company.

The question is why professional portfolio managers would not step in to stabilise the stock market (that is, buy ice firm shares in winter and sell them in summer to take advantage of the ignorant masses who do not understand that the ice business is seasonal). The reason is that professionals are focused on outwitting the general public: "The actual, private object of the most skilled investment today is 'to beat the gun', as the Americans so well express it, to outwit the crowd, and to pass the bad, or depreciating, half-crown to the other fellow" (Keynes 1936: 155).

He went on to describe the modern stock market as similar to a particular kind of beauty contest:

> *[P]rofessional investment may be likened to those newspaper competitions in which the competitors have to pick out the six prettiest faces from a hundred photographs, the prize being awarded to the competitor whose choice most nearly corresponds to the average preferences of the competitors as a whole; so that each competitor has to pick, not those faces which he himself finds prettiest, but those which he thinks likeliest to catch the fancy of the other competitors, all of whom are looking at the problem from the same point of view. It is not a case of choosing those which, to the best of one's judgment, are really the prettiest, nor even those which average opinion genuinely thinks the prettiest. We have reached the third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practise the fourth, fifth and higher degrees. (Keynes 1936: 156)*

Keynes defined this as 'speculation' (trying to anticipate average opinion) rather than "enterprise" (trying to calculate actual business prospects), and argued:

> *Speculators may do no harm as bubbles on a steady stream of enterprise. But the position is serious when enterprise becomes the bubble on a whirlpool of speculation. When the capital development of a country*

> *becomes a byproduct of the activities of a casino, the job is likely to be ill done. The measure of success attained by Wall Street, regarded as an institution of which the proper social purpose is to direct new investment into the most profitable channels in terms of future yield, cannot be claimed as one of the outstanding triumphs of laissez-faire capitalism which is not surprising, if I am right in thinking that the best brains of Wall Street have been in fact directed towards a different object.* (Keynes 1936: 159)

These speculative whirlwinds affect investment in plant and equipment. Euphoric expectations increase a firm's valuation, making it easier to raise funds for new investment; pessimism not only makes it expensive to raise funds, but it also makes it cheaper to just buy a competitor rather than investing in new plant and equipment. For those reasons, speculation in stock markets affects investment, which then drives employment and output.

Keynes provided an investment theory of the cycle in which investment itself depends on long-term expectations. Keynes did not argue that stock prices and investment are always swinging wildly. Recall that in normal times, he believed we safely fall back on conventions. However, the potential for instability always lurks in the background because we know it is unsafe to hold these conventions too strongly. Keynes was sceptical that the 'free market', operating either on the basis of *laissez faire* or with guidance from countercyclical monetary policy, would be able to keep investment on a steady growth path. Instead, he said he expected to:

> *see the State, which is in a position to calculate the marginal efficiency of capital goods on long views and on the basis of the general social advantage, taking an ever greater responsibility for directly organising investment; since it seems likely that the fluctuations in the market estimation of the marginal efficiency of different types of capital, calculated on the principles I have described above, will be too great to be offset by any practicable changes in the rate of interest.* (Keynes, 1936: 164)

Followers of Keynes have developed in more detail the connection between the credit system and economic fluctuations, going beyond Keynes' focus on the stock market. For example, Albert Wojnilower (1980) (who spent a career in financial markets) developed a 'credit crunch' theory of the cycle. In his view, the demand for credit is highly interest inelastic; in other words, the quantity demanded falls very little as rates rise. Especially at an economic cycle peak, the demand for credit is insatiable. Credit typically expands over the course of the cycle. At the peak, however, the quantity of credit can become supply-constrained, facing an insatiable demand that it cannot meet.

Hence, financial conditions have a dominant and powerful effect on the economic cycle: if the supply can expand, it can satisfy demand and allow growth of employment and output. But if there is a sudden imposition of constraint on the supply side, the demand cannot be met. According to Wojnilower, a 'credit crunch' is necessary for a downturn – the downturn in borrowing and spending occurs because suppliers refuse to meet the demand (not because borrowers have decided to reduce borrowing). In the past, the credit crunch has been precipitated either by institutional constraints (such as interest rate limits on deposits) or by defaults by some borrowers (that fuels caution on the part of lenders).

James Crotty (1986) argues that once we add money as a means of payment, with private, legally enforceable contracts, we also have the rise of credit, which is the promise of deferred payment. This extends the separation in time that ordinarily exists in commodity production: instead of C-M (selling a commodity for money), in the first step we have C-D (commodity sold on the basis of credit, with D signifying debt of the buyer), then C-M (the buyer must sell a commodity to obtain money), and finally M-D (the buyer retires the credit advanced by the seller). This opens the possibility of crisis because if the buyers cannot make good on the credits, the seller will have to sell the debts at a discount, which can lead to fire sales, defaults on debt, and runs out of debt/credit and into money.

For this reason, Crotty (1986) argues that credit is a dominating accelerator and destabiliser of accumulation: in an upturn it fuels accumulation, increases profit expectations, and increases debt leverage that can generate a runaway boom. Over the course of the boom interest rates start to rise and illiquidity rises due to leveraging. Then if credit is suddenly withdrawn, or if profits are lower than expected, this leads to a rupture of the system due to interlocking commitments as the chain of payment obligations is broken (if my debtor cannot pay me, I cannot pay my creditor). A crisis results, and is manifested as a run to liquidity, that is, into the safest assets such as treasury bonds, insured deposits, and cash.

As in the case of the Marxist theories of the cycle, all of these Keynesian approaches emphasise credit as a potentially destabilising force.

26.5 Minsky's Financial Instability Hypothesis

Hyman Minsky extended Keynes' analysis (Minsky, 1975; 1986) by developing what he labelled "a financial theory of investment and an investment theory of the cycle", attempting to synthesise the approaches of those who emphasise financial factors and those who stress real factors as causes of the cycle by noting that the two are joined on a firm's balance sheet (Papadimitriou and Wray, 1998).

As in Keynes' approach, fluctuations of investment drive the economic cycle. However, Minsky explicitly examined investment finance in a modern capitalist economy, arguing that each economic unit takes positions in assets (including, but not restricted to, real physical assets) that are expected to generate income flows by issuing liabilities that commit the unit to debt service payment flows.

Because the future income flows cannot be known with certainty (while the schedule of debt payments is more or less known), each economic unit operates with margins of safety, collateral, net worth, and a portfolio of safe, liquid assets to be drawn upon if the future should turn out to be worse than expected. The margins of safety, in turn, are established by custom, experience, and rough rules of thumb. If things go at least as well as expected, these margins of safety will prove in retrospect to have been larger than what was required, leading to revisions of operating rules. Thus, a 'run of good times' in which income flows are more than ample to meet contracted payment commitments will lead to reductions of margins of safety.

Minsky developed a classification scheme for balance sheet positions that adopted increasingly smaller margins of safety:

- Hedge (expected income flows sufficient to meet principal and interest payments). Minsky's use of the term hedge here is not related to 'hedge funds', which typically manage wealth portfolios for high-income individuals, university endowments, and sovereign wealth funds.
- Speculative (near-term expected income flows only sufficient to pay interest).
- Ponzi (expected income flows not even sufficient to pay the interest, hence funds would have to be borrowed merely to pay the interest).

This leads directly to Minsky's most famous contribution, the **financial instability hypothesis**. Over time, the economy naturally evolves from one with a 'robust' financial structure in which hedge positions dominate toward a 'fragile' financial structure dominated by speculative and even Ponzi positions. This transition occurs over the course of an expansion as increasingly risky positions are validated by the booming economy that renders the built-in margins of error superfluous, and encouraging the adoption of riskier positions. Eventually, either financing costs rise or income comes in below expectations, leading to defaults on payment commitments. As in the Marx-Fisher analyses, bankruptcies snowball through the economy. This reduces spending and raises planned margins of safety. The recession proceeds until balance sheets are 'simplified' through defaults and conservative financial practices that reduce debt leverage ratios.

Central to Minsky's exposition is his recognition that the development of the 'big bank' (central bank) and the 'big government' (government spending that is large relative to GDP) helps to moderate cyclical fluctuation. The central bank helps to attenuate defaults and bankruptcies by acting as a lender of last resort; countercyclical budget deficits and surpluses help to stabilise income flows.

While mainstream economists focused on the supposed ability of central banks to control the money supply, or to control the economy through adjustments to interest rates, in Minsky's view the most important function of the central bank is to stop bank runs and prevent 'fire sales' of assets. If the central bank stands ready to lend reserves on demand in lender of last resort operations, it can stop bank runs. Further, by lending to financial institutions so that they can buy financial assets, or through direct purchases of assets by the central bank, the central bank can set a floor to asset prices. This prevents the Fisher-type debt deflation process that was so devastating during the Great Depression of the 1930s.

At the same time, the 'big government' fiscal policy (net spending) can set a floor to aggregate demand, constraining the depths of the downturn, so that a recession will not turn into another Great Depression. Minsky identified three effects of the countercyclical movement of the government's fiscal position toward deficits or larger deficits in a downturn:

- The **income and employment effect**: As government spending rises and taxes fall in a downturn, this helps to maintain private sector income and employment. This is the 'Keynesian' multiplier impact. However, Minsky put more emphasis on the automatic stabilisers than on discretionary 'fiscal stimulus' spending. The automatic stabilisers include transfer payments (welfare, food stamps, unemployment compensation) that rise automatically when the private sector sheds jobs. Tax revenue automatically falls as private sector income falls, and if the tax system is highly progressive the impact is enhanced. Finally, Minsky also included government interest payments on its rising debt (due to deficits) because these provide income to bondholders in the private sector.
- The **cash flow effect**: In a downturn, government payments to the private sector help to provide the 'cash' needed to make payments, including payments due to debt commitments. In his later work, Minsky referred explicitly to the Kalecki profits equation (see Chapter 25): government deficits add to profits, maintaining the cash flows that firms need to service debt.
- The **portfolio effect**: Government deficits in a downturn provide safe and liquid assets to the private sector. As we know from our exposition in Chapter 14, government deficits create net financial wealth in the private sector, accumulated in the safest form (cash, reserves, and treasury bonds).

Hence, a 'big government' fiscal policy can be a powerful stabilising tool, with much of the stabilisation occurring automatically. That can be supplemented by a discretionary fiscal stimulus, taking the form of tax relief or new spending, if the automatic stabilisers are not sufficient to offset the non-government sector spending withdrawal.

The problem, according to Minsky, is that successful stabilisation through the big bank and the big government creates moral hazard problems because economic units will build into their expectations the supposition that intervention will prevent another major depression from happening again. Thus, risk taking is rewarded and systemic fragility grows through time, increasing the frequency and severity of financial crises even when depression is avoided. While there may be no ultimate solution, Minsky believed that informed and evolving regulation and supervision of financial markets is a necessary complement to 'big bank' and 'big government' intervention. In other words, if the government is going to 'backstop' the private sector in bad times, it must regulate financial markets to prevent excessive risk taking in good times.

Conclusion

This chapter examined business cycles in more detail, addressing different kinds of cycles and discussing several different heterodox theories of the cycle, including Marxian, Keynesian, Post-Keynesian and Minsky's financial instability hypothesis. We also explained why orthodoxy has no plausible theory of crises because the economy is presumed to be guided by an invisible hand toward market equilibrium.

When the GFC hit major economies around the world, many observers in academia, and policymaking circles turned to Minsky's theory to try to understand what went wrong. Like Keynes, Minsky dismissed the belief that reliance upon an invisible hand would eliminate financial instability. Indeed, he was convinced that an unregulated, small government capitalist economy would be prone to great depressions and the sort of debt deflation process analysed by Irving Fisher in the 1930s. While a 'big government' capitalist economy could attenuate the downturns, that would make the economy subject to explosive booms that would inevitably go bust, and the bigger the boom, the more difficult it would be to manage the bust.

Minsky was particularly prescient when he argued that if the government successfully used fiscal and monetary policy to rescue the economy after each financial crisis, this would change behaviour in such a way that the financial system would become even more fragile. This is the downside to the "big government" (spending greater than 20% of the economy) and "big bank" (central bank lender of last resort operations) system that has been ready to intervene in crises since the Second World War. This is why it is particularly important that the 'big government/big bank' backstop be accompanied by tight financial regulation. Unfortunately, Minsky's wisdom was ignored during the run up to the GFC as orthodox economists wrongly endorsed deregulation of banks. This is a topic to which we turn in Chapters 31 and 32.

References

Baran, P.A. and Sweezy, P.M. (1966) *Monopoly Capital: An Essay on the American Economic and Social Order*, New York: Monthly Review Press.

Bernanke, B.S. (2004) "The Great Moderation", Board of Governors of the US Federal Reserve System, at the meetings of the Eastern Economic Association, Washington DC, 20 February.

Crotty, J.R. (1986) "Marx, Keynes and Minsky of the Instability of the Capitalist Growth Process and the Nature of Government Economic Policy", in D. Bramhall and S. Helburn (eds), *Marx, Keynes and Schumpeter: A Centenary Celebration of Dissent*, Armonk, NY: M.E. Sharpe, pp 297–326.

Fisher, I. (1933) "The Debt-Deflation Theory of Great Depressions", *Econometrica*, 1 October, 337–57.

Galbraith, J.K. (1972) *The Great Crash*, Boston, MA: Houghton-Mifflin.

Gillman, J.M. (1957) *The Falling Rate of Profit, Marx's Law, and its Significance to Twentieth Century Capitalism*, London: Dennis Dobson.

Kalecki, M. (1943) "Political Aspects of Full Employment", *Political Quarterly*, October, 14(4), 322–30.

Kalecki, M. (1954) *Theory of Economic Dynamics: An Essay on Cyclical and Long-Run Changes in Capitalist Economy*, New York: Rinehart.

Keynes, J.M. (1936) *The General Theory of Employment, Interest, and Money*, New York and London: Harcourt Brace Jovanovich, 1957 Reprint.

Kindleberger, C. (1989) *Manias, Panics, and Crashes: A History of Financial Crises*, New York: Basic Books.

Marx, K. (1990) Capital: Volume 1, Penguin Classics: London.

Marx, K. (1991) Capital: Volume 2, Penguin Classics: London.

Marx, K. (1992) Capital: Volume 3, Penguin Classics: London.

Minsky, H. (1975) John Maynard Keynes, New York: Columbia University Press.

Minsky, H.P. (1982) *Can "It" Happen Again: Essays on Instability and Finance*, New York: M.E. Sharpe.

Minsky, H. (1986) *Stabilising an Unstable Economy*, New Haven, CT and London: Yale University Press.

Papadimitriou, D. and Wray, L.R. (1998) "The Economic Contributions of Hyman Minsky: Varieties of Capitalism and Institutional Reform", *Review of Political Economy*, 10, 199–225.

Sherman, H. (1991) *The Business Cycle, Growth and Crisis under Capitalism*, Princeton, NJ: Princeton University Press.

Steindl, J. (1952) *Maturity and Stagnation in American Capitalism*, New York and London: Monthly Review Press.

Veblen, T. (1904) *The Theory of the Business Enterprise*, New Brunswick, NJ: Transaction Books.

Wojnilower, A. (1980) "The Central Role of Credit Crunches in Recent Financial History", *Brookings Papers on Economic Activity*, 277–326.

Wolfson, M. (1994) *Financial Crises: Understanding the Postwar U.S. Experience*, Armonk, NY and London: M.E. Sharpe.

PART G

HISTORY OF MACROECONOMIC THOUGHT

CHAPTER

27 OVERVIEW OF THE HISTORY OF ECONOMIC THOUGHT

Chapter Outline

Learning Objectives

- Understand the distinctive features of the different schools of economic thought.
- Recognise the importance of the dominant school of economic thought for the design of policy.

27.1 Introduction

It is neither desirable nor possible to present a detailed history of economic thought in this text – that is a topic for another course in economics. However, it is useful for students to have a general overview of the history of economic thought in order to understand the origins of the modern schools of thought (which we will divide into one of the two main approaches to economics discussed in Chapter 1: neoclassical or heterodox). We will trace the evolution of economic thought from the 'father of economics', Adam Smith, to the present.

27.2 History of Neoclassical Theory

We've already mentioned Smith's notion of an 'invisible hand' that guides self-interested market participants to act in their own interests as well as those of society as a whole. Over time, this concept developed into the proposition that if economic 'agents' (those operating in markets, including households and firms) take prices as signals, the market can generate an equilibrium set of prices to clear all markets.

It turns out that this is a very difficult thing to prove. The first major attempts to do so were made by William Jevons, Léon Walras and Carl Menger, all in the early 1870s. While Jevons was most transparent in his effort to prove that capitalism is best for all, Walras was the most rigorous in his attempt. J.B. Clark (writing in the 1890s) expanded the mission to go beyond showing that 'free market capitalism' is 'most efficient' to include an attempt to prove that it is 'fair'. For that task, he developed marginal productivity theory, which can be summarised as showing that the allocation of resources is fair since each person gets rewarded based on their contribution to the production process.

By 1890 this neoclassical framework was well-established, even if it had not entirely succeeded in its main tasks.

This effort continued into the late 1950s when rigorous models based on new mathematical techniques were developed for the purpose. The most important contributions were made by Arrow, Hahn and Debreu, who carefully identified the conditions necessary to demonstrate the existence of a 'general equilibrium' (a vector of relative prices to clear all markets – the rigorous neoclassical definition of equilibrium). Unfortunately, over the following years it was shown that while, under very strict conditions, the existence of a general equilibrium could be proven, the market-clearing vector of relative prices is not unique and it is not stable. What this means (in very simple terms) is that we cannot say whether one equilibrium is better than another, nor can we claim that, if the system is not in equilibrium, the invisible hand of free markets would push the economy toward equilibrium.

While it might not be obvious, both of these negative results are devastating. Essentially, the first says that there are many, perhaps an infinite number of, possible equilibria. Without introducing some additional criteria, we cannot assess whether one equilibrium is better or worse than another. The initial distribution of resources influences the equilibrium outcome, so we could redistribute initial endowments and get a different general equilibrium, but as policymakers we cannot say that one distribution is better than any other. We could begin with an equal distribution, or perhaps we begin with one person owning almost everything. There would be a different general equilibrium in each case, neither of which can be said to be better than the other.

The other negative result – that the market system does not send signals that move the economy to equilibrium – undermines the hypothesis about the operation of the invisible hand. If the economy is out of equilibrium, the resulting price signals are likely to move it further away. Perhaps we need the 'visible hand' of government, after all, to move the economy to equilibrium.

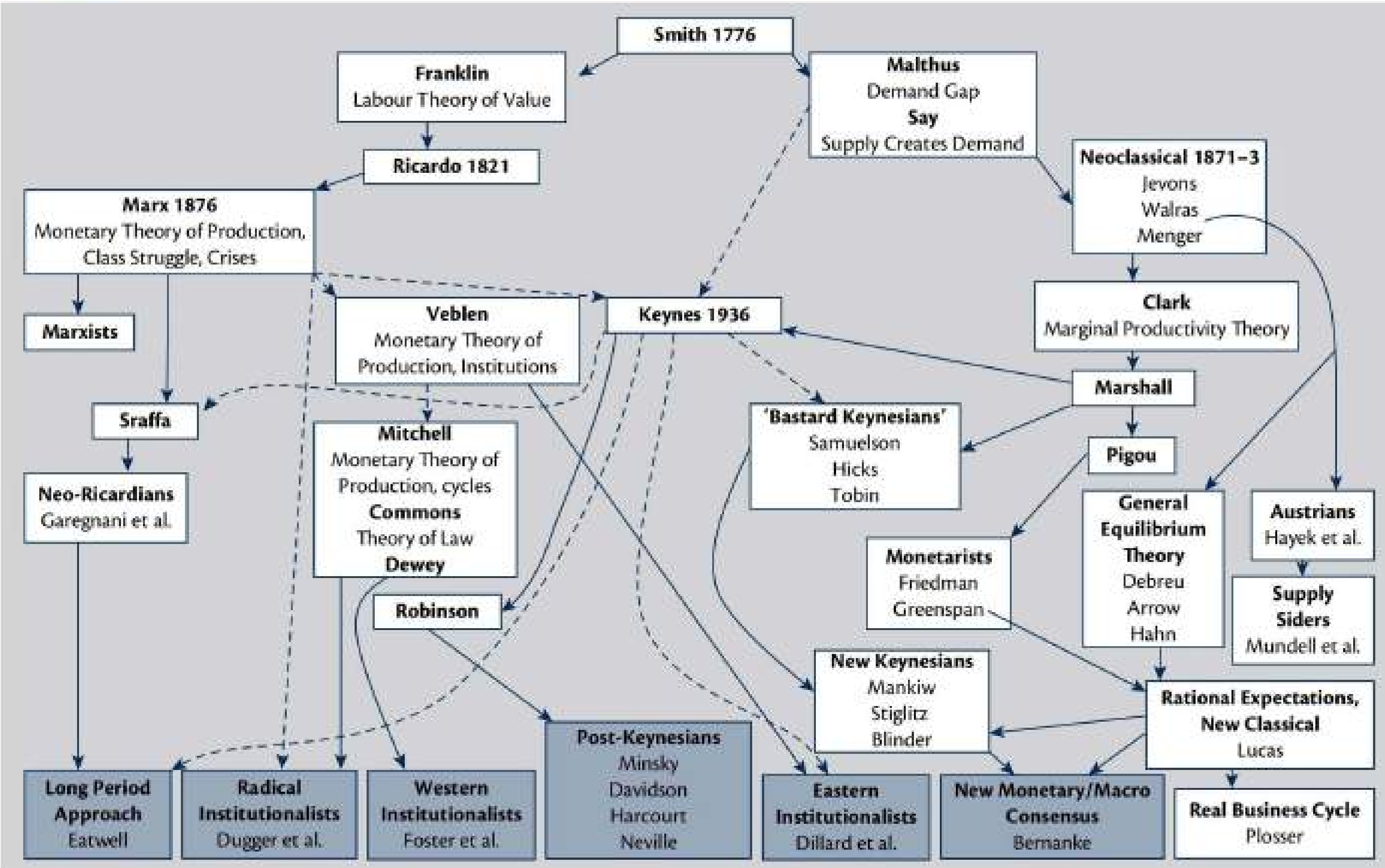
The largely unsuccessful effort to demonstrate the superiority of the market system must be viewed in the context of Karl Marx's masterpiece, *Das Kapital* (or *Capital*), which was published in 1867. (As an indicator of its influence, Jevons made it clear that his tract (1911) was meant to be a response to Marx's work.) Marx's book made two main points. First, capitalism is based on exploitation. Labour produces all value but capitalists are able to appropriate surplus labour power for their own use. In an important sense this is not fair; it is not that labour should receive all of the value that it produces, but rather that surplus labour ought to be at the disposal of society as a whole, rather than appropriated by a few to benefit themselves.

Second, Marx explained that history is driven by class struggle, by those who produce trying to reduce appropriation by the upper classes: slaves against slave owners; peasants against feudal lords; and workers against capitalists. Capitalism actually strengthens the efforts of the working class by bringing them together in factories where they can organise resistance. Revolts against the upper classes (especially against capitalists) will gain momentum until finally workers overthrow their oppressors and institute a new kind of society, one in which workers are on top, and capitalists and other members of the upper classes are oppressed. This is called Socialism, which will eventually evolve into Communism, a new form of society in which class society is banished. In a communist society, everyone is a worker who enjoys contributing to the benefit of society as a whole.

We do not need to assess the accuracy of Marx's analysis here. What is important is to recognise that the mainstream economists needed to respond to Marx, and neoclassical economics was developed to do so. Let us quickly look at the evolution of neoclassical economics after the 1870s. We will not spend a lot of time here discussing the modern neoclassical schools of thought, as they will be treated separately in Chapter 29 *Modern Schools of Economic Thought*.

Figure 27.1 shows the chronology of macroeconomics from Smith to the present. Moving down the right-hand branch, the Austrian school developed from the work of Menger and Walras. The main feature of their approach is the devotion to the supposed ability of the invisible hand of free markets to achieve equilibrium. Of note, however, is the Austrian acceptance of the importance of decision making in historical time in conditions of uncertainty. Unlike the most rigorous neoclassical theory, which allows only for 'logical time' and 'risk', Austrians

Figure 27.1 Chronology of macroeconomics



(like heterodox economists) admit that our expectations about the unknowable future can be wrong. Our mistakes disrupt the movement of the economic system toward equilibrium and cause business cycles. However, the market still 'knows best' so government should not intervene to try to offset the cycle. Indeed, government intervention is more likely to make matters worse. According to Austrians such as Hayek, it is best to let recessions and depressions run their course because they cleanse the system of imbalances.

Continuing down the right-hand side of Figure 27.1, supply siders share the Austrians' distrust of government. Gaining prominence first in the early 1970s, and becoming prevalent during the time in office of President Reagan and Prime Minister Thatcher, they insisted that the best way to get the economy out of the doldrums was to reduce taxes on the rich and eliminate government regulations. These actions would encourage 'the supply side' (the rich) to save and invest more. The Supply Siders are best known for two propositions: 'trickle down' and the 'Laffer curve'. The first is the notion that policies to benefit the rich (such as tax cuts) would actually help the poor as jobs 'trickle down' to them. The second was developed by Laffer, who argued that government budget deficits could be reduced by cutting tax rates, especially on the rich, because this would unleash the entrepreneurial spirit, causing the economy to grow so fast that tax revenues would grow enough to balance the budget. Both of these policies were put in place under President Reagan in the USA. There is now a broad consensus that neither worked: the budget deficit grew after the tax cuts, and inequality increased at a faster pace (and continued to rise up to and through the Global Financial Crisis (GFC)).

To the left of the Austrians in our chronology we have the branch that leads to the modern neoclassical schools of thought: General Equilibrium (GE), Monetarist, New Classical (NC), and Real Business Cycle (RBC) theories. We briefly discussed the GE approach above, the modern representative of which is called Dynamic Stochastic General Equilibrium (DSGE). The DSGE model combines elements from the modern New Classical and New Keynesian approaches into a highly mathematical framework which requires too much technical proficiency to be treated formally in this textbook. The Monetarist approach derives from Marshall (Keynes' professor), Pigou, and Robertson (the latter two were sparring partners of Keynes), and is based on Neoclassical theory.

Monetarism is most closely associated with the work of Milton Friedman (1968a: 98) who claimed that "inflation is always and everywhere a monetary phenomenon", and who attributed the Great Depression to mistakes made by the Federal Reserve. He argued that when the central bank creates too much money, we get inflation. When it creates too little we get a recession or depression. The best course of action then is for the central bank to target a constant rate of growth of the money supply (say, four per cent per year). Otherwise, there is little role for the government to play. Like most neoclassical economists, Monetarists want to constrain government to a few essential areas such as military defence and perhaps the justice system. We discussed Monetarism at length in Chapter 18.

The New Classicals (most prominently Lucas) added to Monetarism the proposition that all economic actors hold 'rational expectations' and that markets clear instantaneously. We'll discuss these notions in detail later, but for now it is sufficient to understand that the New Classical (NC) version of Monetarism is simply a more extreme version of neoclassical economics, essentially carrying it to its logical, if somewhat absurd, extreme. The idea is that since markets always clear, there really is no such thing as unemployment. The only exception is when the central bank 'fools' workers and employers by acting differently to expectation, but given the assumption of rational expectations, such 'fooling' occurs only randomly and temporarily. Markets will quickly right themselves. Hence, during the decade-long Great Depression of the 1930s there really was no problem of unemployment. Rather, people without jobs were voluntarily taking a leave of absence in line with their utility maximisation behaviour. Thus, at the going real wage, they preferred leisure over work.

Real Business Cycle theory presumes that money is always neutral, so that temporary 'fooling' cannot occur. What appears to be a 'business cycle' is not a cycle at all, but rather represents a 'random walk' that results from random shocks to the economy. The most important of these are shocks to productivity, which are likely to occur due to large and random technology shocks that increase or reduce labour productivity. When a negative shock hits the economy, real wages fall with reductions to labour productivity; utility maximising workers leave their jobs because leisure offers more utility than the lower real wage. A positive shock raises productivity and real wages, inducing maximising individuals to take jobs. Thus, there really is no such thing as involuntary unemployment because everyone is always optimising and the economy is always in equilibrium. This is not to say that those who are not working would not be happier with higher wages and jobs, but that given their labour productivity and the real wage they could get from working, they prefer leisure.

The last orthodox school comprises the Bastard Keynesians and the New Keynesians. On the eve of the GFC of 2008, the dominant orthodox schools of thought adhered to the Real Business Cycle theory and the New Keynesian theory. Before discussing these models, however, we will address the heterodox alternative.

27.3 History of Heterodox Thought

Moving down the left-hand side of Figure 27.1, the school of thought from Smith through David Ricardo to Marx is called the Classical school.[1] The two central features of this school are its class-based approach and its labour theory of value. As discussed above, the two main classes under capitalism are the capitalists and the workers who struggle over the distribution of output. Workers do not own the means of production (tools, machinery, factories) so generally must work for capitalists. Capitalists prefer to maintain a 'reserve army of the unemployed' to hold down wages. If their workers demand higher wages, capitalists can threaten to recruit from among the ranks of the unemployed. There is no tendency toward equilibrium at full employment; rather the economy is subject to periodic crises that might become increasingly severe.

According to the labour theory of value, labour produces all value, but receives only a portion of the value it has produced. The rest goes to capitalists as profits (and to other classes such as feudal lords, government officials, and religious authorities as rent, taxes, tithes, and other deductions from the surplus value created by workers). Equilibrium is not defined as a position of market clearing, but rather as a tendency toward equal rates of profit on capital.[2] Today's Marxists carry on in this tradition.

In the centre of our chronology we have John Maynard Keynes. He and Karl Marx are arguably the two key figures in the development of modern macroeconomics. Keynes' influences were Thomas Malthus (a contemporary of Ricardo and a vehement critic of the Classical approach) as well as his own teacher, Alfred Marshall. Note that both of these have been included within the neoclassical tradition.

From Malthus, Keynes adopted the theory of insufficient aggregate demand. Briefly, that is the recognition that while it must be true that aggregate income generated from production is great enough to purchase aggregate output (which is the foundation of the neoclassical Say's Law: 'supply creates its own demand'), those who receive income might not spend it. This was the basis of one of the greatest debates in economic literature: a discourse between Ricardo and Malthus, with Ricardo holding to Say's Law and Malthus arguing that we need an 'unproductive' class that consumes without producing in order to fill 'demand gaps'.

To be sure, Marx also had a theory of 'effective demand', and it is probable that Keynes was also influenced by Marx's theory. Indeed, Marx's theory of demand was more coherent than the argument of Malthus (and quite similar to Keynes' theory) because Marx recognised that there are two different kinds of spending (consumption and investment) while the Ricardo-Malthus debate was solely about consumption.

Keynes argued that Say's Law is broken because people tend to spend only a fraction of their incomes on consumption, which opens a 'demand gap' that must be filled by another kind of spending, investment. When investment is too low, we have a problem of 'effective demand' and thus there are unemployed resources, including labour. For Keynes, this is a normal situation, and hence in general Say's Law does not hold.

From Marshall, Keynes learned an appreciation for including institutional detail. Like Ricardo, Keynes was intimately involved in financial markets. He made and lost, and remade, fortunes in the stock market, commodities markets, and foreign exchange markets. He was also a top advisor to the UK Treasury, and was already famous for a number of books written prior to his *General Theory* (GT) (1936) that developed 'Keynesian Theory', and modern macroeconomics itself as a separate discipline from microeconomics. (Keynes would not have been happy about that because his macroeconomics was fully integrated with his microeconomics.)

Keynes was a prolific writer on a variety of topics, and many of his books were influential. He wrote a treatise on the Indian currency and another on probability theory. His widely acclaimed critique of the reparations imposed on Germany after the First World War (*The Economic Consequences of the Peace*, 1920) correctly predicted that the burden would prove to be too much to bear. This contributed to the Weimar hyperinflation as well as to the conflicts that helped pave the way to Hitler's rise to power. He also wrote major works on money and monetary policy.

Keynes authored an interesting little pamphlet, "The End of *Laissez-Faire*" (1926) in which he presented a powerful rejection of the notion of an invisible hand. In some ways that piece laid the groundwork for his GT. At that time he did not present an alternative to neoclassical economics, although he knew what was wrong with it. In 1930 he began to develop an alternative theory in the two volumes of *The Treatise on Money*, but he wrote to friends immediately on publication that he was already dissatisfied with it. He had not yet 'escaped' neoclassical theory. Finally, in 1936 he provided a coherent alternative that in many ways still serves as the basis for the macroeconomic approach we present in this textbook.

The Great Depression showed that Keynes' intuition back in 1926 had been correct. The invisible hand not only did not move the economy back to full employment equilibrium, but in fact made things worse. As sales fell, firms laid off workers. When workers lost their jobs and income, they cut back on purchases. Falling sales led to more lay-offs, and to the slashing of wages and prices. Unable to cover their costs, firms went out of business, leading to more job losses.

Rather than falling wages and prices generating more hiring and more sales, they led to precisely the opposite. Left to its own devices, the invisible hand might have driven unemployment higher in the 1930s. Government had to intervene to stop the decline into ever-deeper depression.

Keynes' GT provided the theoretical framework to explain what needed to be done, and how it would work to stop the downward spiral. To be sure, the GT provides much more than a justification for government intervention, but the evidence of the depression gave Keynes' book much credence. It very quickly became a classic.

There was a small group of (younger) economists working with Keynes at Cambridge University, known as the 'Circus', reading and commenting on drafts. After Keynes' death in 1946, they carried on the Keynesian tradition, mostly in the UK and in Europe (especially in Italy) and Australia. 'Keynesianism' also spread to the USA, propagated by Alvin Hansen and Paul Samuelson. However, in America Keynes' macroeconomics ideas were 'synthesised' with neoclassical theory to produce what Samuelson called the 'Neoclassical Synthesis'. J.R. Hicks had developed the IS-LM model in 1937 (see Chapter 28) and this served as the basis for American Keynesianism.

The Circus followers of Keynes, especially Joan Robinson, rejected the American version of Keynesianism and proceeded to develop and extend Keynes in a manner that they considered to be closer to the true spirit of his GT. This came to be called Post-Keynesian economics. The best-known early Post-Keynesians are Paul Davidson and Hyman P. Minsky (USA), John Neville and Geoffrey Harcourt (Australia), Thomas Rymes (Canada), and Pierangelo Garegnani, Luigi Pasinetti, and Paolo Sylos-Labini (Italy).

Over time, several distinct strands of Post-Keynesian theory were developed, including the Neo-Ricardians (followers of Piero Sraffa, who worked with Keynes and who edited the works of Ricardo), the Surplus or Long Period approach (which combined Sraffian and Keynesian economics), the Franco-Italian Circuitist approach (among whom Augusto Graziani is one of the best known), the 'Fundamentalist Keynesian' approach (Davidson), and the 'Financial Keynesian' approach of Minsky.

There are also Post-Keynesians who focus on developing an alternative microeconomics that is consistent with Keynesian macroeconomics. Many of these follow the work of Michał Kalecki, the Polish economist who was at Cambridge with Keynes. Among other contributions, Kalecki developed the 'mark-up' approach to pricing, as well as the 'Kalecki profits identity' (discussed in Chapter 25), and explored business cycle dynamics.

27.4 Institutional Economics

The final major heterodox figure we need to discuss is Thorstein Veblen, an American social scientist who came up with a number of terms that we still use, such as the 'leisure class', 'invidious distinction', 'conspicuous consumption', and 'pecuniary emulation' (also called 'keeping up with the Jones's' in some countries). He is the founder of the Institutionalist approach to economics that emphasises the role of culture, norms of behaviour, and institutions. By institution Veblen meant much more than say, a financial institution or a religious institution. He included patterns of behaviour (for example, gift-giving at Christmas) and even patterns of thought (pecuniary emulation). He posited a dichotomy between the 'industrial arts' (producing goods and services) and the 'business enterprise' (making monetary profits).

Like Keynes, Veblen rejected the neoclassical belief that the invisible hand of the market would guide everyone to serve the public interest. Indeed, in 1919 he wrote a short pamphlet titled "On the Nature and Uses of Sabotage", arguing that the "captains of industry" (capitalists) had created cartels to sabotage (conspire to reduce) output in order to keep prices up. He predicted that this would lead to a major economic depression, which turned out to be quite correct. In many respects, Veblen's analysis of business cycles was similar to Keynes' theory.

Veblen's most important analyses of capitalism were *The Theory of Business Enterprise* (1904) and *Absentee Ownership and Business Enterprise in Recent Times: The Case of America* (1923). However, his most popular book was *The Theory of the Leisure Class* (1899). While in many respects Keynes was an 'insider', holding influential positions in academia and policymaking, Veblen was the ultimate 'outsider' who skewered his profession (see *The Higher Learning In America: A Memorandum On the Conduct of Universities by Business Men*, 1918) as well as 'high society'.

Veblen's followers included John R. Commons (who developed theories of law and economics) and Wesley Mitchell (the 'father' of business cycle theory), and later John Kenneth Galbraith (perhaps the most famous American economist). Several branches of Institutional economics emerged, as shown in Figure 27.1. While some of the earlier Institutional economists were somewhat wary of Keynes' economics, most of the later Institutionalists have integrated the approaches of Keynes and Veblen.

The Radical Institutionalists also draw heavily on Marx. A point of departure from Marx however, is that many Institutionalists follow Charles Peirce, John Dewey and Clarence Ayres in adopting 'instrumental value theory' and 'pragmatism' rather than Marx's labour theory of value. While these are difficult philosophical concepts, Institutionalists emphasise that the purpose of theory (indeed, of thought) is as an instrument to take action to solve problems. What is important is the choice of the means to reach an end. Human thought and action are purposive, and should be judged on the basis of their ability to achieve the purpose.

The final point to make is that most of the heterodox schools of thought follow the three main figures, Marx, Veblen, and Keynes, in adopting a 'monetary theory of production', or what Veblen (1904) called a "theory of business enterprise". In a capitalist economy, the purpose of production, at least that which is undertaken by capitalists, is to make profits. To undertake production generally requires money to buy or hire the inputs to the production process. The output is then sold in markets for money. The hope, of course, is that the capitalist ends up with profits, which requires that the output is sold for more money than it cost to produce. As Marx put it, we can view the process as M-C-P-C'-M' where we start with some amount of money (M) to buy commodity inputs (C); P is the production process that transforms C to final commodities (C') for sale to realise money (M', which includes profits). So long as M' is greater than M, profits have been made (M' – M = Profit) (see Chapter 26).

27.5 Modern Orthodox Schools of Thought

Note that we have drawn a dotted line in Figure 27.1 from Keynes to modern 'Bastard Keynesians', the version of Keynes developed largely by American 'Keynesians' such as Samuelson, James Tobin, and Robert Solow. The use of the word 'bastard' comes from Robinson, who proclaimed that while we know who is the 'mother' of this approach (that is, neoclassical economics), we really do not know who the father is. In her view it was not Keynes.

Still, these Neoclassical Synthesis Keynesians (as Samuelson called them) claim allegiance to Keynes, and do accept the theory of effective demand. While they believe that market forces will eventually move the economy in the direction of market-clearing equilibrium, they argue that this could take too long, causing unnecessary suffering among the unemployed.

Hence, they advocate use of monetary and fiscal policy to 'fine tune' the economy, hastening the move to equilibrium (essentially, by assisting the sometimes weak 'invisible hand' of markets through judicious use of government intervention). They claim that there is a trade-off, since trying to lower unemployment will increase inflation (the Philips Curve trade-off that we addressed in Chapter 18).

In the 1980s an offspring of Neoclassical Synthesis Keynesianism was developed, called the 'New Keynesian' approach. It adopted much of the framework of the New Classical economics of Lucas, including rational expectations. However, advocates of this approach added various kinds of 'frictions' to the market that prevent it from moving quickly to equilibrium.

One friction was already well known, namely minimum wage laws that were alleged to prevent wages from falling sufficiently to clear the labour market. New Keynesians incorporated many other frictions that prevent wages, prices, and interest rates from moving quickly and by large enough amounts to restore equilibrium. Together these frictions lead to 'sticky wages and prices'. Gradually, the notion that wages and prices are sticky gained popularity and many mainstream models now explicitly adopt this assumption in their models and claim that this inflexibility impedes the adjustment of the economy to market-clearing equilibrium.

The dominant mainstream approach is now the New Monetary Consensus which includes many of the features of these recent neoclassical schools of thought: sticky wages and prices, rational expectations, a Philips Curve unemployment-inflation trade-off, an aggregate demand equation (similar to the investment saving curve in the IS-LM model, which we introduce in Chapter 28), and a role for monetary policy (a replacement for the money demand-money supply curve in the IS-LM model). Often monetary policy is based on a 'Taylor Rule' which was developed by John Taylor in 1993. The central bank sets an interest rate target in response to 'output gaps' but also with the objective of keeping inflation on target.

This has necessarily been a very brief introduction to the history of economic thought. We will discuss many of these modern economic schools of thought in more detail in Chapters 28 and 29. The background provided here is merely designed to place these modern schools into historical context, so we can see how they are linked to the greatest economists since the days of Adam Smith. It is no substitute for a serious investigation of the history of economic thought, which can only be provided in a course devoted to the topic. You might want to refer back to the economic chronology in Figure 27.1 when we discuss particular schools of thought in more detail.

27.6 Post-War Economic History and History of Thought

Let us finish this chapter with a survey of the links between post-war economic history and the development of economic thought over the period. We conclude with the implications of the GFC for mainstream theory, and for the possibility of a return to Keynes' own economics.

When the Great Depression began after September 1929, the schools of economic theory were not clearly defined. The most commonly used textbook was Marshall's *Principles of Economics* (1890[1930]), which used diagrams to introduce supply and demand, the marginalist approach, and utility theory. In the text, Marshall provided institutional and historical detail to illuminate the arguments. He was the most important contributor to the rise of British 'Political Economy' based on neoclassical theory.

Marshall's student, John Maynard Keynes, became editor of the top British academic journal, the *Economics Journal*, at a young age. However, when it came time to criticise neoclassical theory in his *General Theory* in 1936, Keynes had to systematise a still somewhat incoherent approach. He used another of Marshall's students, Arthur Cecil Pigou, as something of a foil, arguing that in order for Pigou to reach the conclusion that falling wages would resolve the unemployment problem, he must believe certain propositions. (For example, see Keynes' *General Theory*, Chapter 2.)

On the eve of the Second World War, the links between theories and policy recommendations were not yet coherent. In other words, there was nothing like Samuelson's (1958) post-war textbook, *Economics: An Introductory Analysis* to lay out the 'principles' in a systematic manner.

Further, 'Classical' British Political Economy was not widely accepted in the USA. Indeed, the top US journal was the *American Economic Review*, founded in opposition to the neoclassical approach by American Institutionalists. Economists working in the Institutionalist tradition played a big role in economic policy, embarking on policy experiments, especially at the level of US states, studies that would prove useful in the 1930s. American policy was pragmatic, so when the Depression hit, even the conservative President Hoover began to implement 'New Deal'-type policies. Policymakers did not wait for economic theory to catch up to real world events; they knew they had to do something.

However, they did not do enough, and it was not until President Roosevelt took office in 1933 that more decisive action was taken. This included a 'bank holiday' and fundamental reform of the financial system; creation of the New Deal jobs programmes that employed a total of 13 million workers; and a huge growth in the size of the Federal government, which took over responsibility for building the nation's infrastructure and for guiding the economy. Roosevelt brought into his administration many of the nation's Institutionalist economists, who would continue to play a big role in policy formation all the way to the 1960s (for example, Galbraith served presidents from Roosevelt to Kennedy).

However, as discussed above, 'Keynesianism' gradually took over economics in the UK, the US and elsewhere. Institutionalists were at first suspicious. In the hands of economists such as Samuelson, Keynesianism was simplified and rendered 'mechanical', with simple macroeconomic models that downplayed Institutionalist concerns with institutional details. Questions about distribution, financial institutions, sectoral impacts of policy, and even national industrial and employment policy were shunted to the side as economic policymakers emphasised 'pump priming', that is, general aggregate demand stimulus. They argued that growth alone would resolve most problems because 'a rising tide lifts all boats'. By the end of the 1960s, Samuelson was even quoted as claiming that recessions and inflations were a thing of the past, because economists knew how to fine-tune the economy.

But the economists were mistaken. Recessions soon returned in the USA in the 1970s, and inflation had already accelerated in many countries by the end of the 1960s. Because mainstream economics had endorsed the Philips curve trade-off, it was not possible to fight the simultaneous appearance of inflation and unemployment – what would later be called stagflation. Problems grew throughout the 1970s as economic growth slowed and stagflation persisted.

This allowed neoclassical economic theory to return with a vengeance, in a more radical form than that against which Keynes had fought. Milton Friedman and his followers had been developing Monetarism as an alternative to 'Keynesianism' throughout the 1950s and 1960s, and the events of the 1970s gave them the opportunity to strike. Friedman had never accepted the fine-tuning argument nor the Philips curve trade-off. His 1968 American Economic Association Presidential Address (Friedman, 1968b) had presciently predicted *accelerating* inflation (meaning that he foresaw no stable trade-off of inflation for unemployment). His policy recommendations were radically 'free market': he opposed social security for the aged, subsidised health care (Medicare) for the sick, and even viewed the licensing of doctors as unnecessary government interference in private decision-making. However, with the rise of neoliberalism (also called neoconservatism in the USA) from the time of US President Reagan and UK Prime Minister Margaret Thatcher, these radical recommendations became mainstream.

Over time, many of Friedman's claims and policy recommendations were called into question, even by the mainstream. Federal Reserve Chairman Paul Volcker tried to implement Friedman's 'money growth rate rule' – fix the rate of growth of the money supply at some constant but low rate, to fight inflation in 1979. By 1982 Volcker gave up because the Fed couldn't hit the target and in any case, the growth of the money supply and the rate of inflation did not appear to be systematically related. (The Bank of England had a similar experience with money targeting at that time.) By the late 1980s the US Fed gave up any pretence of controlling the money supply and by the 1990s no major central bank was still targeting the money supply because all of them switched to the 'Keynesian' recommendation to target the overnight interest rate.

Further, it became apparent that money supply growth and nominal GDP (or national income) growth are not as closely linked as Friedman had always claimed, meaning that even if government could control the growth of the money supply, that would not allow it to control inflation. Friedman's claim that inflation could be brought down quickly and with little economic cost (in terms of rising unemployment) through proper monetary policy, also seemed to be inconsistent with real world experience. For example, inflation was reduced during the 1980s, but only at the cost of lower economic growth and much higher unemployment rates.

In the sphere of economic theory, neoclassical economists moved to rational expectations and rejected Friedman's theory that central bank policy works by 'fooling' workers and employers (discussed in Chapter 18). Many rejected the distinction between the 'short run' where monetary policy is said to be potent (hence, money is 'non-neutral'), and the 'long run' where it is impotent (money is 'neutral'), in favour of the presumption that anticipated monetary policy is always neutral. The mainstream moved on to Real Business Cycle theory, where (apparently) cyclical swings are due to 'real factors' such as fluctuations of labour productivity. The recommended role for government to play in the economy shrank far below that prescribed even by Friedman, as neoclassical theory 'circled the wagons' against virtually any government intrusion into the economy. The invisible hand metaphor became more powerful than it had ever been.

The GFC of 2008 led to "Keynes: The Return of the Master", as Keynes' biographer, Robert Skidelsky, titled his 2009 book. It called into question the whole edifice of the neoclassical dogma that the invisible hand knows best, and led to growing interest in heterodox economists such as Marx, Veblen, Keynes, and especially Minsky, who had developed a theory of financial crises based on Keynes' insights.

Skidelsky nicely skewers the rational expectations revolution in that book: "Rarely in history can such powerful minds have devoted themselves to such strange ideas" and also criticises New Keynesians: "Having swallowed the elephant of rational expectations, they strained at the gnat of the continuous full employment implied by it, and developed theories of market failure to allow a role for government" (2009: xiv–xv). He goes on to argue that the efficient markets hypothesis "has been the biggest casualty of the current financial meltdown" (p. 38). In short, the problem with all versions of orthodoxy is the "mistaken belief that all risk can be correctly priced and that therefore financial markets are optimally self-regulating" (p. 51).

Skidelsky provides an excellent overview of the innovations, delusions, and events that preceded the crisis, and explains how Keynes' insights help us to analyse this crisis and to formulate a conditional policy response.

What is particularly interesting is that Skidelsky links all of this to Keynes' ethics, morals, and early beliefs. For Keynes, 'good' behaviour is not a self-seeking 'rational' (in the neoclassical sense) endeavour, rather it is, following the philosopher G.E. Moore, action that contributes to "good states of mind". In Keynes' view, good is objective, and "[r]ational people know what is good" (Skidelsky, 2009: 136–7). Economics, as a moral philosophy, is about how one ought to behave, and how to "improve the quality of our desires to the point that they become desirable" (Skidelsky, 2009: 138).

According to Keynes, "To make the world ethically better was the only justifiable purpose of economic striving" (Skidelsky, 2009: 133). The goal is not to maximise pleasure (what neoclassical economics calls utility) but rather goodness; this is easier to do once individuals "have a certain level of material comfort" (p. 139).

The problem is "love of money" ("the root of all evil") because it is morally inefficient (the objectless pursuit of wealth) and economically inefficient (the psychology of the miser depresses demand) (Skidelsky 2009: 142). The conflict for Keynes as a moral philosopher is that the love of money drives capitalism, which helps to provide the required material comfort necessary to live "wisely and agreeable, and well" (p. 144), but it is also a neurosis because there can never be a point of satisfaction (p. 153).

Orthodoxy largely ignores this 'neurotic' love of money because it is irrational and cannot be measured or quantified, yet the mainstream approaches promote unbridled wealth accumulation and often even accept the concentration of wealth and income that follow on from so-called 'market' outcomes that are taken as given. The social, macro, implications are unexamined because of the methodological approach that aggregates up from individuals (in extreme versions, from a representative agent). The implications of this are taken up in the *General Theory's* Chapter 24: the fetish for liquidity plus inequality generates chronic demand gaps and unemployment.

Skidelsky (2009) relates Keynes' analysis to the crisis at the time: financialisation (or what Minsky called 'money manager capitalism') of the past several decades had elevated the love of money: today "wealth increase is the only goal western society has to offer… Because its only goal is to make societies richer, capitalism has to be more successful than any rival economic system in order to survive" (p. 134). When the Berlin Wall fell and the 'free market' was embraced globally, it appeared that capitalism was triumphant. Yet, according to Skidelsky, it destroyed communities, wrecked the environment, generated record inequality, even fuelled wars and terrorism, before it collapsed into a global crisis.

Skidelsky (2009) argues that the crisis highlights three failures: an institutional failure ("banks mutated from utilities into casinos"), an intellectual failure (economists advocated a dangerous view of the efficacy of markets), and a moral failure ("a system built on money values", "the worship of economic growth for its own sake") (pp.168–9).

Still, Skidelsky (2009) predicted that we would not experience a 1930s-style depression in the aftermath of the financial crash of 2008 because we have Keynes. Governments around the world instinctively returned to the Master; fiscal stimulus "is one of the bequests of the Keynesian revolution" (p. 19), which was supported by inter-governmental organisations, including the OECD, IMF, World Bank and even the European Union. In large part, we owe the initial policy response to the GFC to the return of "the most brilliant mind of modern times" (p. 192).

However, by 2011, the OECD and IMF were pushing fiscal consolidation measures to address "adverse deficit/debt dynamics", albeit with some qualifications (Sharpe and Watts, 2012). Under the rules of the Maastricht Treaty, Eurozone countries had little capacity to enact fiscal stimulus measures anyway, and monetary policy was based on a 'one size fits all' model with the European Central Bank setting a common target interbank rate (see Chapter 23). While US President Obama had pushed for an economic stimulus, it lasted only two years before deficit hawks insisted on austerity. Around the world, a dangerous new doctrine of 'growth through austerity' became a mantra, driving fiscal tightening. As a result, the global recovery languished in most regions (China was an exception). A decade after the onset of the GFC, some countries still had not fully recovered.

Thus, for a while it had looked like orthodoxy was dead, that economics teachers would abandon the neoclassical textbooks and that policymakers would spurn the advice of their neoclassical economists. While that view was overly optimistic, the crisis created a lot of doubt about neoclassical theory and renewed interest in heterodoxy. It became obvious that the orthodox policy prescriptions failed to respond adequately to the crisis.

With the weak recovery in the countries with sovereign currencies, and continuing crisis in some of the European countries that had given up sovereignty to adopt the euro, there was return to the Keynesian notion of a positive government spending multiplier that could boost growth.

Conclusion

Chapter 27 provided a brief introduction to the history of thought. We contrasted the orthodox and heterodox approaches. Although each of these is made up of a number of different schools of thought, there are sufficient similarities that we can identify two quite different approaches to economics in a chronology that stretches back more than 200 years to Adam Smith's *Wealth of Nations* (1776). In spite of the differences among the orthodox approaches, all of them rely heavily on a market-clearing equilibrium approach – even if they do allow some room for 'market failures' that prevent immediate movement to equilibrium. Further, the focus of orthodox approaches is on the 'real' economy in which money is supposed to play at best a minor role. It is no surprise that orthodox economists did not see the GFC coming.

By contrast, heterodox approaches emphasize the peculiar nature of economies that rely on money. While some heterodox approaches do use a notion of equilibrium, it is quite different from the orthodox vision of market clearing. For example, we saw earlier that Keynes' theory of effective demand envisions equilibrium as a 'state of rest' – one in which firms hire the number of workers that is consistent with expected sales, but is not necessarily consistent with full employment. Heterodox economists also include more real world institutional detail in their analyses. As we saw in Chapter 26, they warn that the natural processes of the modern capitalist system generate instability and crises. As such, it was not surprising that many heterodox economists also did see the crisis coming.

Going forward, there is a chance that heterodoxy could regain some of its lost dominance because there is popular revulsion against an orthodox economics that is out of touch with the real world. Only time will tell if there will be such a significant reaction against both an economic system that rewards a single-minded 'love of money' and an economic orthodoxy that celebrates selfishness as the highest moral imperative.

References

Friedman, M. (1968a) *Dollars and Deficits: Inflation, Monetary Policy and the Balance of Payments*, Englewood Cliffs, NJ: Prentice-Hall, Northwestern University.

Friedman, M. (1968b) "The Role of Monetary Policy", Presidential Address to American Economics Association, *American Economic Review*, 58(1), March, 1–17.

Jevons, W.S. (1911) *Theory of Political Economy*, London: Macmillan.

Keynes, J.M. (1920) *The Economic Consequences of the Peace*, New York: Harcourt Brace.

Keynes, J.M. (1926) "The End of *Laissez Faire*", London: Hogarth Press. 1–54.

Keynes, J.M. (1930) *A Treatise on Money, Vol 1, The Pure Theory of Money*, London: Macmillan.

Keynes, J.M. (1930) *A Treatise on Money, Vol 2, The Applied Theory of Money*, London: Macmillan.

Keynes, J.M. (1936) *The General Theory of Employment, Interest and Money*, London: Macmillan.

Marshall, A. (1890[1930]) *Principles of Economics: An Introductory Volume*, 8th edn, New York: Macmillan.

Marx, K. (1867) *Das Kapital*, Berlin: Verlag von Otto Meisner.

Samuelson, P. (1958) *Economics: An Introductory Analysis*, New York: McGraw-Hill.

Sharpe, T. and Watts, M.J. (2012) "Policy Advice in Crisis: How Inter-Governmental Organisations have responded to the GFC", *Journal of Australian Political Economy*, 69, Winter, 103–33.

Skidelsky, R. (2009) *Keynes: The Return of the Master*, London: Allen Lane.

Taylor, J.B. (1993) "Discretion versus Policy Rules in Practice", *Carnegie-Rochester Conference Series on Public Policy*, 39, 195–214.

Veblen, T. (1899) *The Theory of the Leisure Class*, New York: Macmillan.

Veblen, T. (1904) *The Theory of Business Enterprise*, New York: Charles Scribner's Sons.

Veblen, T. (1918) *The Higher Learning In America: A Memorandum On the Conduct of Universities by Business Men*, New York: B.W. Huebsch.

Veblen, T. (1919) "On the Nature of and Uses of Sabotage", *Dial*, LXVI, 341–6.

Veblen, T. (1923) *Absentee Ownership and Business Enterprise in Recent Times: The Case of America*, New York: B.W. Huebsch.

Endnotes

1. The Classical school should not be confused with the neoclassical tradition, which is the main orthodox approach. Unfortunately, Keynes and many others frequently refer to the neoclassical tradition that was attacked by Keynes as the *classical* approach. Further, the orthodox New Classical approach that was developed in the early 1970s has nothing in common with the Classical approach that culminated with Marx's theory. Unfortunately, there is no easy way around this confusing terminology. In this chapter, however, when we use the term 'classical' we are referring only to the tradition that runs from Smith to Marx.
2. American students will be interested to know that Benjamin Franklin contributed to the development of the labour theory of value, as well as being one of America's most influential Founding Fathers.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

28 THE IS-LM FRAMEWORK

Chapter Outline

Learning Objectives

- Be able to derive and interpret the LM and IS schedules and the IS-LM equilibrium.
- Analyse policy options within the IS-LM framework.
- Recognise the macroeconomics consequences of wage/price flexibility.
- Understand the limitations of the IS-LM framework.

28.1 Introduction and the Concept of General Equilibrium

Soon after the publication of Keynes' *The General Theory of Employment, Interest and Money* (GT), British economist J.R. Hicks published an article that attempted to integrate the insights that he felt were useful in the GT with those of the so-called 'Classics' that Keynes had opposed. We discussed the debate between Keynes and the Classics in Chapter 12.

REMINDER BOX

Recall that what Keynes called the 'Classical' approach is actually the orthodox neoclassical approach. The IS-LM approach is not related to the Classical economics discussed in Chapter 27 that we traced from Smith to Marx.

The so-called neoclassical synthesis that emerged and dominated macroeconomic thinking, particularly the textbook expositions, was built on the work of Hicks and his IS-LM model (see Box 28.1 (p. 459) for a more detailed discussion of the history around this).

The IS and LM curves are the equilibrium relationships pertaining, respectively, to the product (goods) market, where investment equals saving, and the money market where the demand for money (liquidity preference) equals the money supply.

The representation of the goods market equilibrium in terms of the simple equality between investment and saving reflected the historical period when the model was developed. It was thought that we would gain essential insights into income determination by assuming a closed economy without government. The IS equation was subsequently extended to allow for the impact of government spending and taxation and net exports on aggregate demand.

The income-expenditure model that we considered in Chapter 15 allowed the interest rate to impact on investment and hence aggregate demand and output in the goods market. The interest rate was considered to be determined by the equality of money demand and supply in the money market. It was treated as a given so that goods market equilibrium could be established.

The IS-LM approach was built on the interdependence of the two markets. The general IS-LM approach showed how the equilibrium solution for output (employment) and the interest rate was simultaneous. In other words, equilibrium in both the goods and money markets would be determined simultaneously. In an analytical sense, there were two unknowns, output and the interest rate, and two equations to solve for these unknowns.

The IS-LM model is an early example of a general equilibrium model. It is still included in most macroeconomic texts and remains fashionable in some mainstream economics circles. In particular, during the fallout following the Global Financial Crisis (GFC), it was used by some well-known economists, including Paul Krugman, who claim that it is still an appropriate approach to adopt.

As we will discuss, the IS-LM model is fundamentally flawed. Indeed, its creator, Hicks, later admitted that it is incoherent. It is highly misleading when used to understand the economy, and dangerous if used to formulate policy. Even orthodox macroeconomists have dropped the model in favour of the New Monetary Consensus and Dynamic Stochastic General Equilibrium models. We are including it in this text only because so many economists and policymakers persist in using it. However, we do not derive the full AD-AS (aggregate demand-aggregate supply) model that flows from the IS-LM approach and is the main organising framework for the orthodox exposition of macroeconomics. That model really adds very little to the IS-LM framework.

28.2 The Money Market: Demand, Supply and Equilibrium

As we noted in Chapter 9, money is a stock and economists have a number of definitions of money, depending on the required degree of liquidity. We wish to consider the factors influencing the desired holdings of notes and coins in circulation plus current bank deposits by the private non-bank sector, that is, M1. In what follows, we follow orthodox practice and use the term *money* to indicate cash plus bank demand (also called current) deposits.

Three motives have been identified for holding money in preference to other assets:

- **Transactions motive**: people need money to engage in daily transactions. Thus, the demand for liquid assets will be a proportion of total national income.
- **Precautionary motive**: additional money balances are held to meet unforeseen events such as urgent car repairs. This motive also tends to vary positively with national income as the higher is the level of economic activity, the higher are the overall transactions.

- **Speculative motive**: Keynes' contribution, which we considered in Chapter 12, was to highlight that money is not simply a means of exchange. People use money in times of uncertainty over movements in interest rates. They have a choice between holding money which earns no interest return or purchasing an interest bearing asset which has less liquidity. Since the GFC, a low interest rate has been paid on current (formerly cheque) deposits, but this fact does not undermine the propositions being advanced in the text.

Keynes juxtaposed the decision to hold money or bonds. If the interest rate is expected to rise, the price of bonds would fall and capital losses would be expected. Thus, at lower interest rates more people would prefer to hold money than take a chance that they would lose should they invest in bonds (whose price would fall if rates rose). Alternatively, if the interest rate falls, the price of bonds rises and capital gains would be enjoyed. At higher interest rates, more people will form the view that rates will fall rather than rise and their liquidity preference would be lower.

The other way of thinking about the impact of interest rates is to recognise that the opportunity cost of holding money rises when interest rates are higher because holding money means that the opportunity to hold an interest earning asset, such as a bond, is foregone.

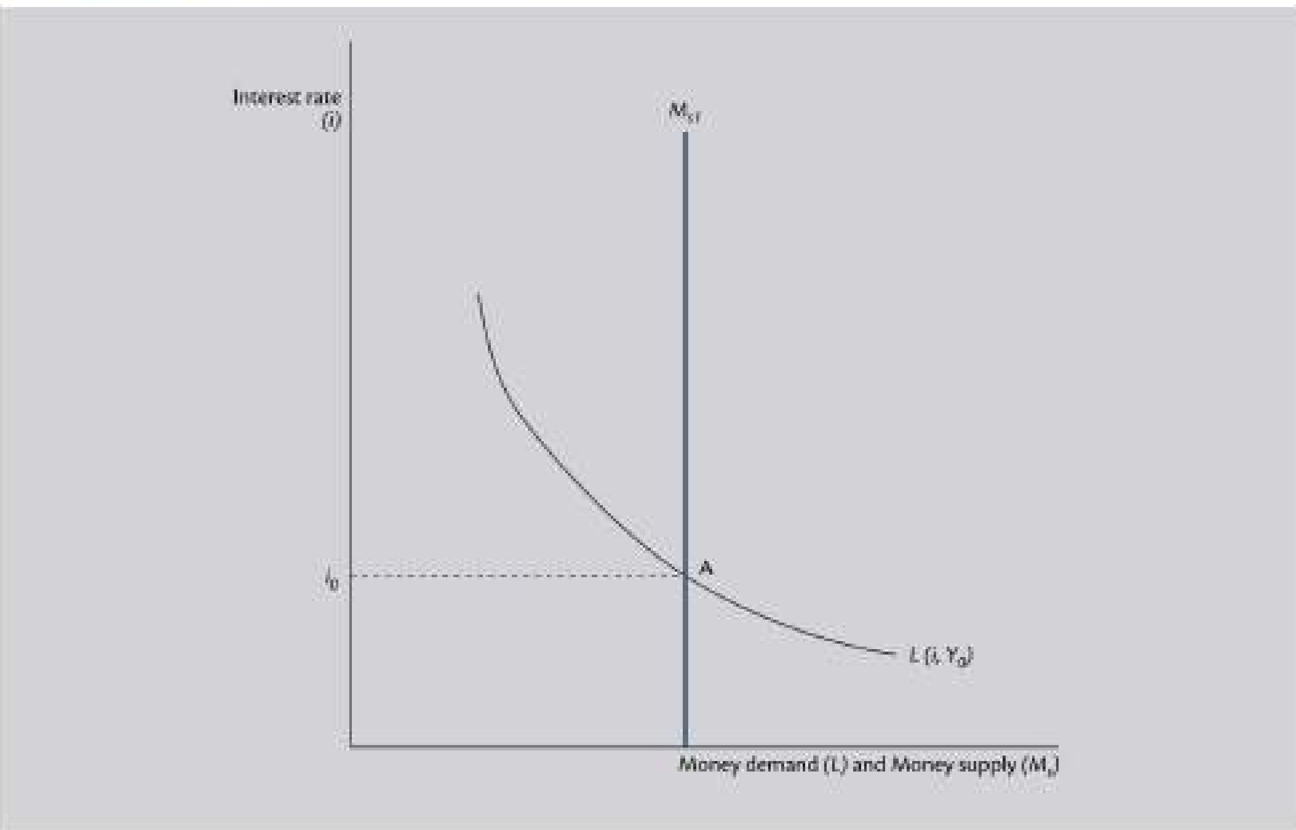
Thus, the demand for money – or as Keynes called it liquidity preference – is a function of both the level of income and the interest rate. Here we are referring to the desire to hold money balances as part of a wealth portfolio.

Figure 28.1 shows the liquidity preference function. The nominal rate of interest is on the vertical axis and the volumes of money demand (L) and money supply (M_s) are on the horizontal axis.

When we draw the money demand curve, we hold income constant. We will assume that the price level is constant. Hence, at this stage, we will treat the level of real income as constant at Y_0. Later we will show that the IS-LM model allows us to introduce price level changes but at present we are operating in real terms.

The money supply (M_s) is assumed to be controlled by the central bank via the monetary base and the money multiplier determines the quantity of money supplied for a given base. Thus, the money supply is treated as exogenous and fixed. We can thus capture that assumption as a vertical line.

Figure 28.1 Equilibrium in the money market



The intersection of the money demand and money supply curves determines the interest rate. Note that in this exposition, the terms money demand and liquidity preference are used interchangeably.

This is consistent with Keynes' departure from the Classical economists who considered the interest rate to mediate saving and investment (that is, the interest rate is conceived to be a real variable, not a monetary variable). Keynes argued that the nominal interest rate is a monetary variable determined by the demand for liquidity and the available supply of money.

Money market equilibrium requires that the demand for money equals the supply.

If the national income level is Y_0 then the intersection of the relevant liquidity preference function, L, and the money supply, M_s, would generate an interest rate of i_0, point A in Figure 28.1.

28.3 Derivation of the LM Curve

The money demand curve (L) is downward sloping with respect to the interest rate and shifts outwards to the right at higher income levels. This is because the demand for money for transactions and precautionary purposes is positively related to the level of income.

Given a fixed money supply, higher income levels are associated with greater money demand and hence higher interest rates.

The higher the interest rate is, the lower will be the quantity of money demanded, other things being equal. However, at any interest rate, the higher the level of national income, the higher will be the demand for money.

The main reason why money demand is a function of the interest rate is because of the speculative motive. As discussed, this relationship is negative because at higher rates, more people expect the rate to eventually fall, which would generate capital gains for bondholders. Thus, at high rates the speculative demand for money is lower; at low rates, the speculative demand is higher.

The money demand curve is smooth and non-linear because of a diversity of opinions about interest rate movements. For example, as interest rates rise, wealth holders progressively form the view that they have reached their maximum. The demand curve would not be a smooth function if everyone held the same expectation. Below we will show what happens when everyone believes that the interest rate has reached the minimum possible (called a liquidity trap).

We have shown that if the national income level is Y_0 then the intersection of the corresponding liquidity preference function, $L(i, Y_0)$, and the money supply line generates an interest rate of i_0, at point A (Figure 28.1).

What would happen if income rose to Y_1, as shown in Figure 28.2?

At the current interest rate i_0 there would now be an excess demand for money (measured by the segment AB) and this would lead to the interest rate rising until the excess demand was eliminated at C, corresponding to interest rate i_1, given that the money supply is treated as fixed.

What would happen if the central bank increased the money supply? Figure 28.3 shows the impact of an increase in the money supply from M_{s1} to M_{s2}.

If the money supply rises to M_{s2} then there is an excess supply of money at i_0 (measured by the segment AB) and the interest rate would drop until it reached i_2 where the demand for money is again equal to the money supply (point D).

Thus, if we start at point A and the national income level rose to Y_1, the interest rate could be held constant at i_0 if the central bank accommodates the increased demand for money by raising the money supply to M_{s2}.

We can now derive the LM curve which shows all combinations of income and interest rates that are consistent with money market equilibrium, for a given money supply.

Figure 28.4 shows the derivation of the LM curve. In the money market diagram (Figure 28.4(a)), points A, B and C represent equilibrium states where money demand equals money supply for different levels of income. As we increase income, we obtain new money demand curves that can be used to calculate the new point of equilibrium in the money market.

Each equilibrium point is thus a unique combination of income and interest rate. We can translate these equilibria to a new graph (as shown in Figure 28.4(b)) where national income (Y) is on the horizontal axis and the interest rate (i) is again on the vertical axis.

Figure 28.2 Money market equilibria

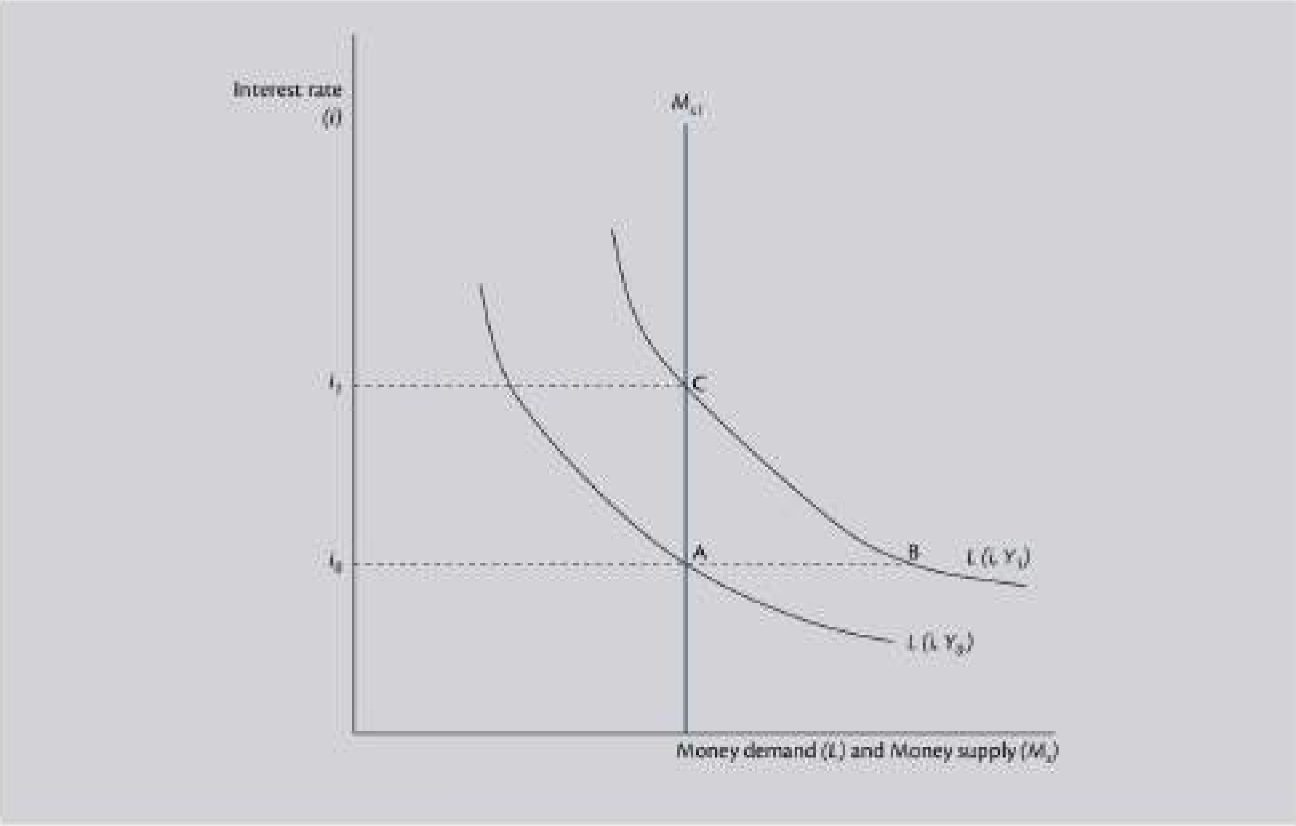


Figure 28.3 The impact of an increase in the money supply from M_{s1} to M_{s2}

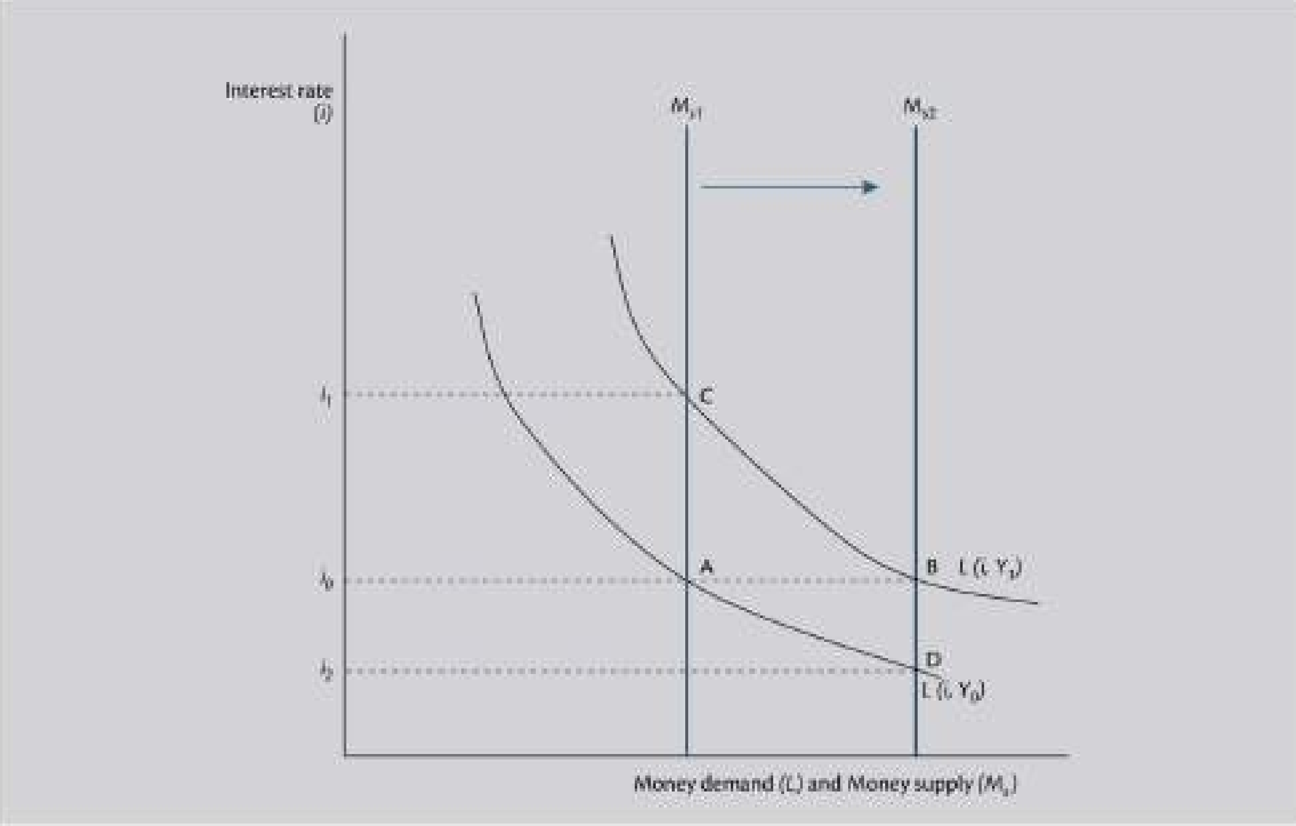
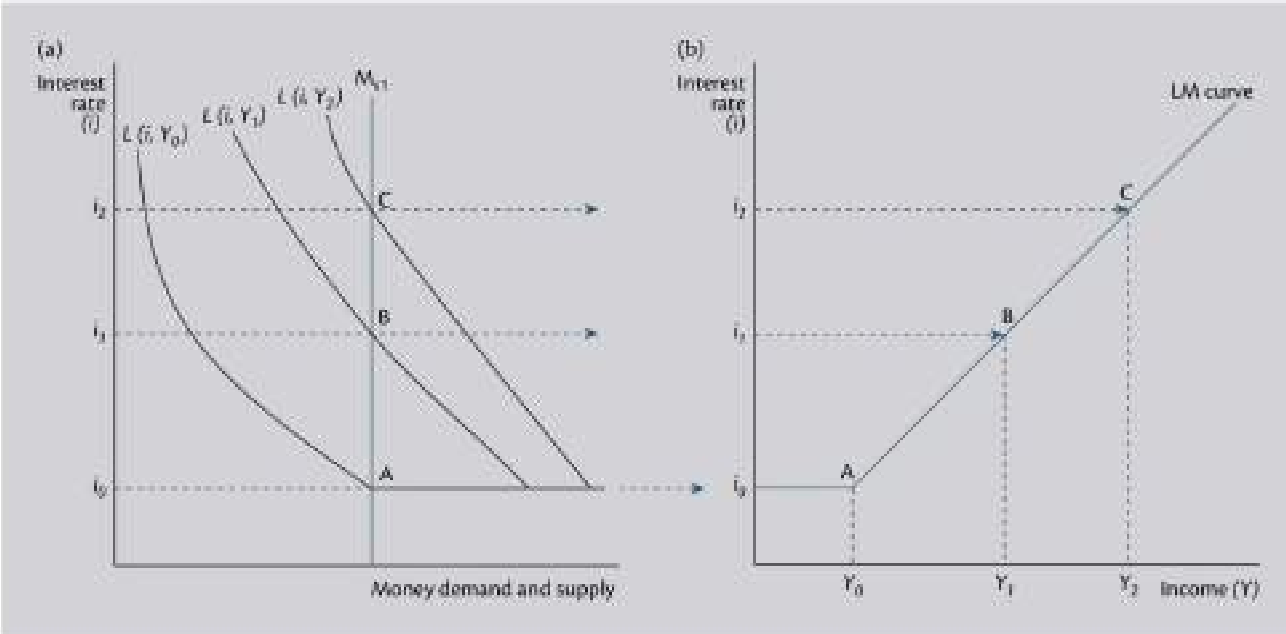


Figure 28.4 The LM curve



Each point in the income-interest space is consistent with money market equilibrium. This set of points is the **LM curve.**[1]

Note that at interest rate, i_0, the LM curve is flat. What does that mean? The horizontal segment of the LM curve relates to the presence of the **liquidity trap.**[2] The liquidity trap arises at some minimum interest rate (which could be zero) where everybody forms the view that the only direction for interest rates is up. The equivalent expectation is that everybody considers that capital losses will be incurred on bond portfolios because when interest rates rise, bond prices fall. The result is that once interest rates reach this minimum level, all people will prefer to hold any new money in the form of cash instead of bonds.

Keynes said (1936: 207):

> *There is the possibility ... that, after the rate of interest has fallen to a certain level, liquidity-preference may become virtually absolute in the sense that almost everyone prefers cash to holding a debt which yields so low a rate of interest. In this event the monetary authority would have lost effective control over the rate of interest ... Moreover, if such a situation were to arise, it would mean that the public authority itself could borrow through the banking system on an unlimited scale at a nominal rate of interest.*

As we will see when we consider policy analysis within the IS-LM framework, the existence of a liquidity trap renders monetary policy ineffective as a stabilisation tool at low interest rates.

Monetary policy is characterised in this IS-LM framework as the central bank manipulating the money supply. When the interest rate is at i_0 in Figure 28.4, increasing the money supply would have no impact on interest rates or the price of bonds. In other words, monetary policy changes cannot alter the level of national income.

In a liquidity trap, a rise in the money supply leads to an equal rise in the demand for money and as a result the interest rate does not change. We will consider this in more detail later in the chapter.

The LM curve is upward sloping at higher levels of income because as national income rises, the demand for money increases and at each given money supply, the interest rate has to rise to ration the excess money demand and maintain money market equilibrium.

The slope of the LM curve is steeper:

- The more sensitive the demand for money (transactions and precautionary motives) is to national income changes. Thus, small increases in national income lead to large changes in excess money demand for a given money supply. The rise in interest rates to restore money market equilibrium, other things equal, has to be larger as a consequence.

- The less sensitive the speculative demand for money is to changes in interest rates. Thus, for a given excess demand for money, the interest rate increase that is required to restore money market equilibrium is larger.

While the horizontal LM curve (liquidity trap case) is one extreme, the other extreme is sometimes referred to as the *Classical case* and describes a vertical LM curve.

The Classical case arises from a demand for money function which is not sensitive to the interest rate. In other words, money is considered to be only a means of exchange and the speculative demand for money is ignored. In this case, the demand for money shifts outwards when income rises and inwards when it falls. As a consequence there is only one national income level consistent with money market equilibrium for a given money supply and the LM curve is vertical.

As an advanced study, in the Appendix to this chapter we derive an analytical solution to the IS-LM framework which shows the impact of these two sensitivities (elasticities) that determine the curve's slope.

Shifts in the LM curve arise from changes in the money supply. Refer back to Figure 28.3, which showed that for a given money demand curve, interest rates fall when the money supply rises. The reasoning is that at a given money market equilibrium combination of interest rates and income, a rise in the money supply generates an excess supply of money, which requires interest rates to fall to stimulate the demand for money sufficiently to absorb the extra supply of money.

In terms of the LM curve, this means that at higher levels of money supply, equilibrium interest rates will be lower at each income level, and this translates into a shift outwards in the LM. The opposite occurs when the money supply falls.

The LM curve can also shift if there is an autonomous change in liquidity preference, which means the money demand rises(falls) at each income level depending on whether the preference for liquidity rises(falls). For example, if people become more pessimistic about the future, they may use increased cash holdings as a haven from uncertainty. This will lead to an outward shift in the money demand curve so that for a given money supply, the equilibrium interest rate will be higher at each income level.

28.4 The Product (Goods) Market: Equilibrium Output

In Chapter 15 we developed the real expenditure model of income determination. From the national accounting framework, we know that total expenditure (E) in the domestic economy in any particular period can be expressed as:

$$(28.1) \quad E = C + I + G + (X - M)$$

Equation (28.1) is identical to Equation (15.1). As it stands, Equation (28.1) is an accounting statement by dint of the definitions and sources of aggregate spending.

The equilibrium level of national income (Y) is determined by aggregate expenditure, such that $Y = E$. The task in Chapter 15 was to understand the behaviour of each of the expenditure components in the national accounting framework and to understand how they interact to determine national income.

We assumed that firms in aggregate plan a fixed volume of investment spending in each period. At that stage of the text we wanted to trace out the implications of changes in autonomous (exogenous) components of expenditure (investment, government, exports and so on) on national income via the multiplier process.

However, in Chapter 25, we developed a more detailed model of investment spending, which allows us to take into account the impact on capital formation of changes in interest rates. So here we assume that total investment spending is influenced, in part, by expectations of future economic conditions and the interest rate, rather than being exogenous, as in Chapter 15.

Business firms continually form expectations about future output. Firms must make resource commitments (for example, working capital, labour) well in advance of realisation (sales), and so the scale of production at any point in time reflects the guesses that they have made in a highly uncertain world.

Further, for given expectations about future sales and revenue, a firm's investment decisions will also be influenced by the cost of capital goods, in addition to the interest rate.

If interest rates rise, the cost of funds necessary to invest in new capital equipment rises and so marginal projects (relative to expected net revenue) may become unprofitable. In other words, investment is likely to be an inverse function of the interest rate, other things being equal.

Therefore, we might hypothesise that total investment is given as:

$$I = b_1 - b_2 i \qquad (28.2)$$

where b_1 is an autonomous component of investment and b_2 is the interest rate sensitivity of investment to interest rate changes.

The higher is b_2, the more investment will decline (rise) for a given interest rate rise (fall).

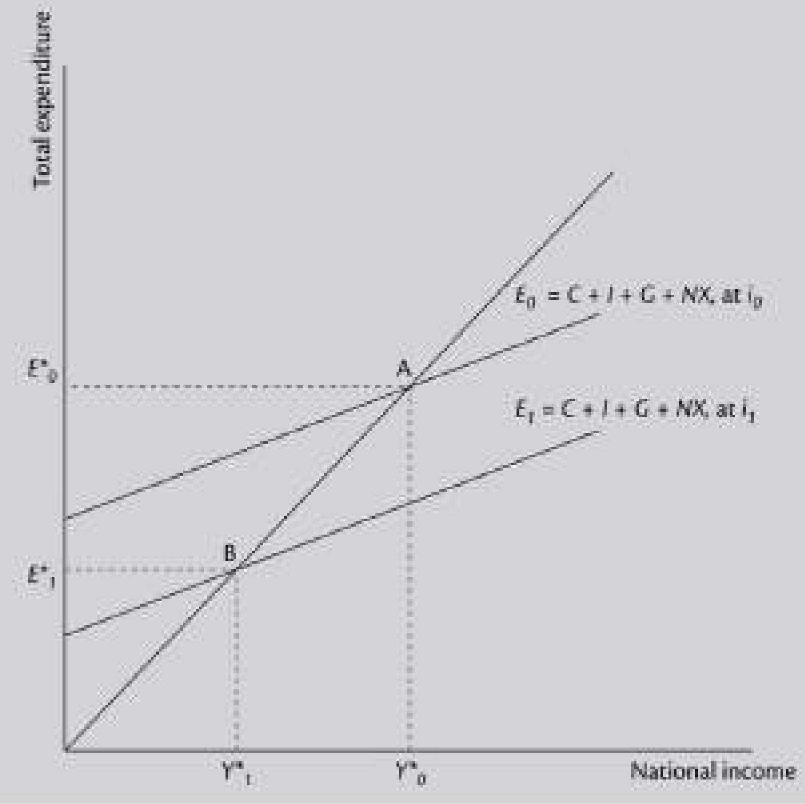
The IS-LM framework retains Keynes' insight that planned savings is a positive function of national income. A more detailed analysis of the GT would also reveal that Keynes considered that the interest rate might also influence consumption spending (via wealth impacts). Further, the purchase of consumer durables such as white goods might require access to consumer credit. However, for now, to keep the argument simple, we assume that the interest rate only impacts on investment.

In Chapter 15 we assumed that firms in the economy are quantity adjusters and so prices are fixed in the short term. Figure 15.6 brought together the 45° aggregate supply curve with the aggregate demand curve (E). It showed that equilibrium national income occurs where the aggregate demand function cuts the 45° line.

At this point, the aggregate demand expectations formed by the firms, which motivate their decisions to supply (Y^*) are consistent with the planned expenditure by consumers, firms, government and the external economy (E^*).

Figure 28.5 augments Figure 15.6 by adding in the impact of Equation (28.2), that is, allowing investment to be inversely impacted by interest rate changes. The total expenditure curve, $E = C + I + G + NX$ is drawn for a given interest rate. The lower the interest rate ($i_0 < i_1$), the higher is investment (and total spending) at all income levels.

Figure 28.5 Product market equilibrium and interest rate changes



Consequently, the total expenditure curve shifts upwards. When interest rates rise, the total expenditure curve would shift downwards, other things being equal.

Point A in Figure 28.5 shows the product market equilibrium associated with an interest rate of i_0. Thus, the combination of income, Y^*_0 and interest rate, i_0 is an equilibrium combination in the product market.

What happens if the interest rate were to rise to i_1? Total investment would decline at all income levels and the total expenditure curve would shift downward from E_0 to E_1.

The excess supply of output at the prior income level leads firms to cut back output and employment and national income falls. A new product market equilibrium occurs when $E^*_1 = Y^*_1$. The combination of income level, Y^*_1, and interest rate level, i_1, also represents an equilibrium combination in the product market.

We thus have two combinations of interest rates and income levels which are consistent with product market equilibrium. Clearly, we could consider the impact of many rates of interest on investment and hence total expenditure and trace out the corresponding equilibrium combinations of interest rates and income.

28.5 Derivation of the IS Curve

The IS curve shows all combinations of interest rates and income at which the product (goods) market is in equilibrium. Figure 28.6 shows this derivation.

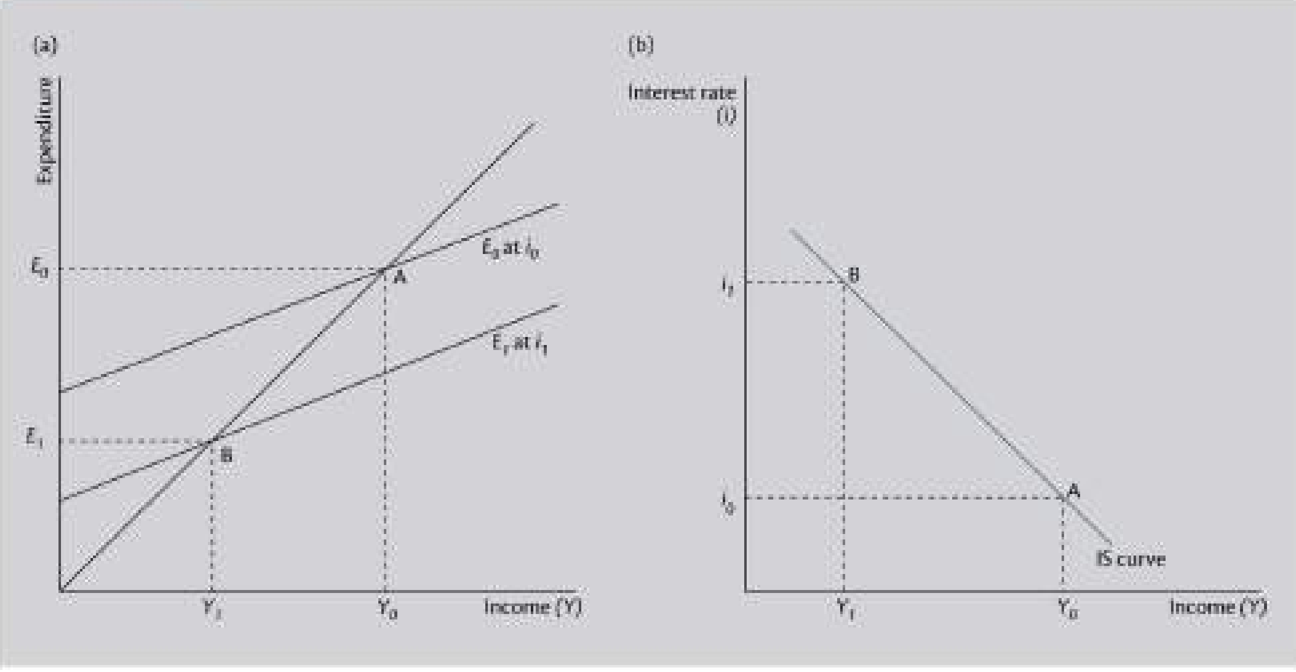
Point A in Figure 28.6(a) is the product market equilibrium where the interest rate is i_0 and total expenditure is E_0, generating total national income of Y_0.

In Figure 28.6(b) where the interest rate is on the vertical axis and national income is on the horizontal axis, point A shows the combination of the interest rate and income which generates the product market equilibrium shown in Figure 28.6(a).

If interest rates rise to i_1, total expenditure falls to E_1 as a result of the lower investment expenditure, which leads to a fall in national income via the expenditure multiplier. Point B shows the new product market equilibrium at (i_1, Y_1).

We could examine the impact of any number of interest rate changes on product market equilibrium in Figure 28.6(a) and subsequently map these points onto Figure 28.6(b). The result would be the IS curve. The IS curve therefore is a series of points representing combinations of national income and interest rates corresponding to equilibrium in the product market.

Figure 28.6 The derivation of the IS curve



In the IS-LM framework, the money market impacts on the product market through the effect of interest rate changes on investment. The change in income results from the initial response of investment to an interest rate change then being multiplied through the expenditure system via induced consumption and leakages to taxation and imports. In other words, the total change in income that follows a change in the interest rate depends on the values of the expenditure multiplier and the sensitivity of investment to interest rate changes.

What factors will shift the IS curve? First, any increase(decrease) in autonomous spending shifts the IS up(down) because for a given interest rate, the equilibrium level of national income rises(falls) when autonomous spending rises(falls).

The magnitude of the shift up(down) in the IS curve resulting from a rise(fall) in autonomous spending is determined by the magnitude of the change in autonomous spending and the size of the expenditure multiplier. For any given change in autonomous spending, the shift in the IS curve will be larger, the greater is the value of the expenditure multiplier.

The slope of the IS curve represents the overall sensitivity of national income to interest rate changes. The larger is the expenditure multiplier, and the more sensitive investment is to interest rate changes, the flatter the IS curve will be. This is because for a given change in interest rates, the effect on investment spending will be larger the more responsive it is to the cost of capital, other things being equal.

In turn, a given change in investment will generate a larger(smaller) change in national income the larger(smaller) is the value of the expenditure multiplier.

If current period investment is very unresponsive to a change in the current interest rate, then the IS curve will be very steep. This is the usual assumption of heterodox economists, and follows on from most empirical work that finds the 'interest elasticity' of non-residential investment to be low.

It is argued, by economists who consider time to be an important consideration in economic analysis, that investment spending plans are based on expectations of future revenue streams that were formed in past periods. The current period's flow of investment spending reflects these past decisions. The time it takes to evaluate different projects, design the appropriate necessary capital equipment, source funding and then build the capital infrastructure suggests that current investment spending will be relatively slow to adapt to current changes in interest rates. We discussed this topic in more detail in Chapter 25.

It should be clear that changes in the tax rate (t), which impact on the value of the expenditure multiplier, will also impact on the slope of the IS curve. For example, a rise in the tax rate will cause the IS curve to become steeper because it reduces the value of the expenditure multiplier due to a larger leakage from the expenditure system.

Similarly, a higher saving propensity or propensity to import, which means that there are larger leakages from the expenditure system each time income changes, will lead to a steeper IS curve.

28.6 Equilibrium and Policy Analysis in the IS-LM Framework

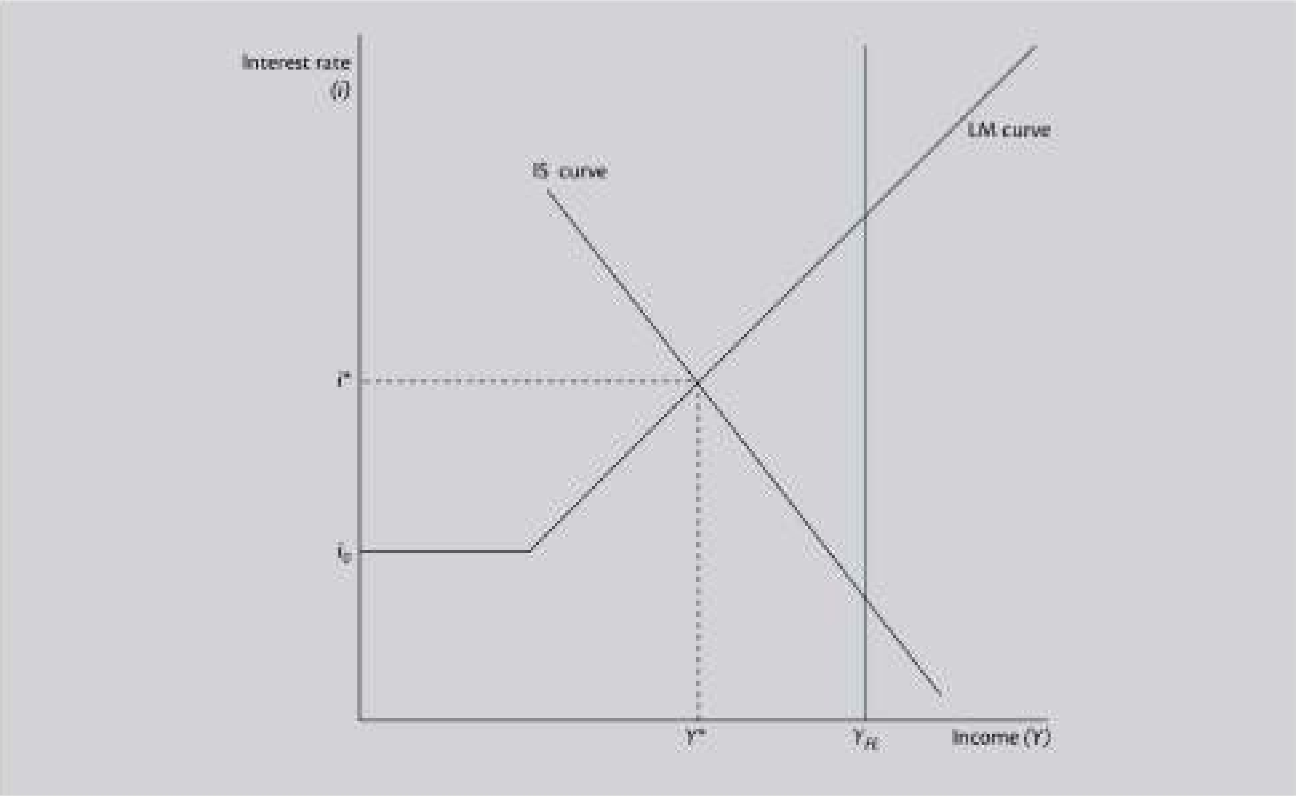
The intention of the IS-LM framework is to combine the product market and money market outcomes into a single diagram so that we can simultaneously determine the equilibrium value of national income and the interest rate. This highlights the interdependency between these markets, a point that Keynes demonstrated clearly. What happens in one market impacts on the other market, which then leads to feedback loops and new equilibrium outcomes in each.

The IS-LM framework thus conceives of a general equilibrium defined as the interest rate and income level that generates simultaneous equilibria in the product and money markets. In a graphical form, this equilibrium position corresponds to the intersection of the IS and LM curves. Figure 28.7 shows the IS-LM solution for equilibrium income and interest rates, Y^* and i^*, respectively.

Two points are worth noting. The vertical blue line at Y_{FE} denotes a hypothetical full employment national output level. In other words, at this output level all available labour and capital are being productively deployed.

The IS-LM joint equilibrium thus can occur at levels of income which are below full employment in the labour market. This is consistent with Keynes' insight that the capitalist system has a tendency to reach steady states at less than full employment, and these need to be addressed by policy interventions.

Figure 28.7 General IS-LM equilibrium



At Y* and i*, business firms are selling as much as they expected to sell and have no incentive to expand production and employment. The desire for liquidity by firms and households is also being fully met by the available supply of money.

This under-full-employment equilibrium can be reached at an interest rate above the minimum rate, where the economy enters a liquidity trap.

The IS-LM framework is used within the mainstream approach to analyse the impact of fiscal and monetary policy changes on output (income) and interest rates, and by implication, employment.

Monetary policy is represented by the assumed capacity of the central bank to alter the money supply. Inherent in this approach is the view that central banks manipulate base money (reserves) which is then transmitted into the broader money supply via the money multiplier mechanism.

REMINDER BOX

In Chapter 10, we demonstrated how this view of central bank operations is not a valid representation of the real world. In fact, the central bank has little control over the money supply and conducts monetary policy principally via its capacity to set the short-term interest rate. However, for the purposes of this chapter, to ensure we render the IS-LM approach faithfully, we assume the money supply is exogenous and under the control of the central bank. Monetary policy changes are thus represented in the IS-LM framework by **shifts** in the LM curve.

Figure 28.3 showed that if the central bank increases the money supply, the interest rate falls at the current national income level. This is because at the existing interest rate, there is an excess supply of money and the interest rate must fall to stimulate an increased demand for money.

The interest rate continues to fall until the demand for money is again equal to the increased money supply and money market equilibrium is restored.

In terms of the LM curve, this means that at higher levels of money supply, equilibrium interest rates will be lower at each income level, which translates into a shift outwards of the LM curve. The opposite occurs when the money supply falls.

The LM curve shifts to the right when the money supply rises and shifts to the left when the money supply contracts.

Figure 28.8 shows the impact of expansionary monetary policy. At some existing monetary policy stance captured by LM_1 the equilibrium combination of the interest rate and national income is i^*_1, Y^*_1. Point A shows the equilibrium position where LM_1 cuts the IS curve.

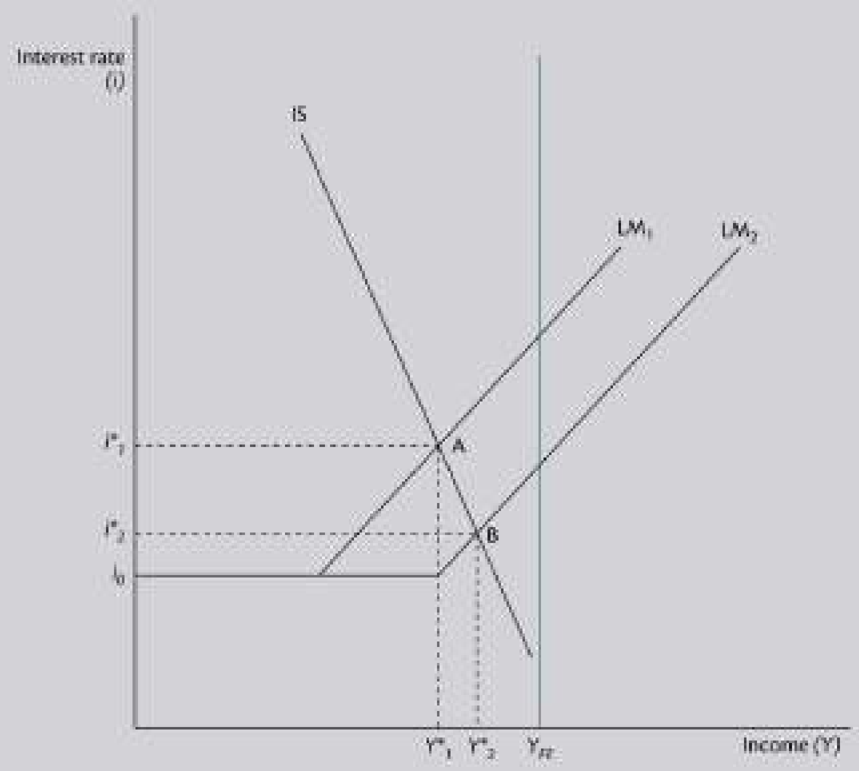
The central bank decides that the output gap, which is measured by the difference between the full employment national income level (Y_{FE}) and the current national income level (Y^*_1), is excessive, given the implied mass unemployment that would be associated with such a deficiency in output. The central bank decides to increase the money supply.

The LM curve shifts to LM_2, which drives down interest rates. The new equilibrium is at point B, with the equilibrium combination of the interest rate and national income at (i^*_2, Y^*_2).

The rising income results from the positive impact on investment of the lower interest rates (represented by the movement along the IS curve from A to B). The more sensitive is investment spending to interest rate changes, the more expansionary the monetary policy change will be.

Note, that in this case, expansionary monetary policy would be unable to achieve full employment because the economy would encounter a liquidity trap (at i_0) before full employment was restored. It is not possible to push the interest rate below i_0 because everyone has come to expect that rates will only go up; in other words, the speculative demand for money becomes infinite so that no one will give up holding money in order to hold bonds. Monetary policy becomes impotent.

Figure 28.8 Expansionary monetary policy



A contractionary monetary policy could be represented in Figure 28.8 by a shift in the LM curve from LM_2 to LM_1. This would drive interest rates up and national income down. The falling income results from the negative impact on investment of the higher interest rates (represented by the movement along the IS curve from B to A). The more sensitive investment spending is to interest rate changes, the more contractionary the monetary policy change would be.

Expansionary monetary policy drives the interest rate down and national income up. Contractionary monetary policy drives the interest rate up and national income down.

The effectiveness of monetary policy depends on the slope of the IS curve. The steeper the IS curve, the smaller will be the impact of monetary policy changes on equilibrium national income. Monetary policy also becomes ineffective in a liquidity trap.

Fiscal policy changes could be implemented by discretionary changes in government spending or the tax rate. We have learned that a rise in government spending shifts the IS curve to the right because for a given interest rate, the equilibrium level of national income rises when autonomous spending rises.

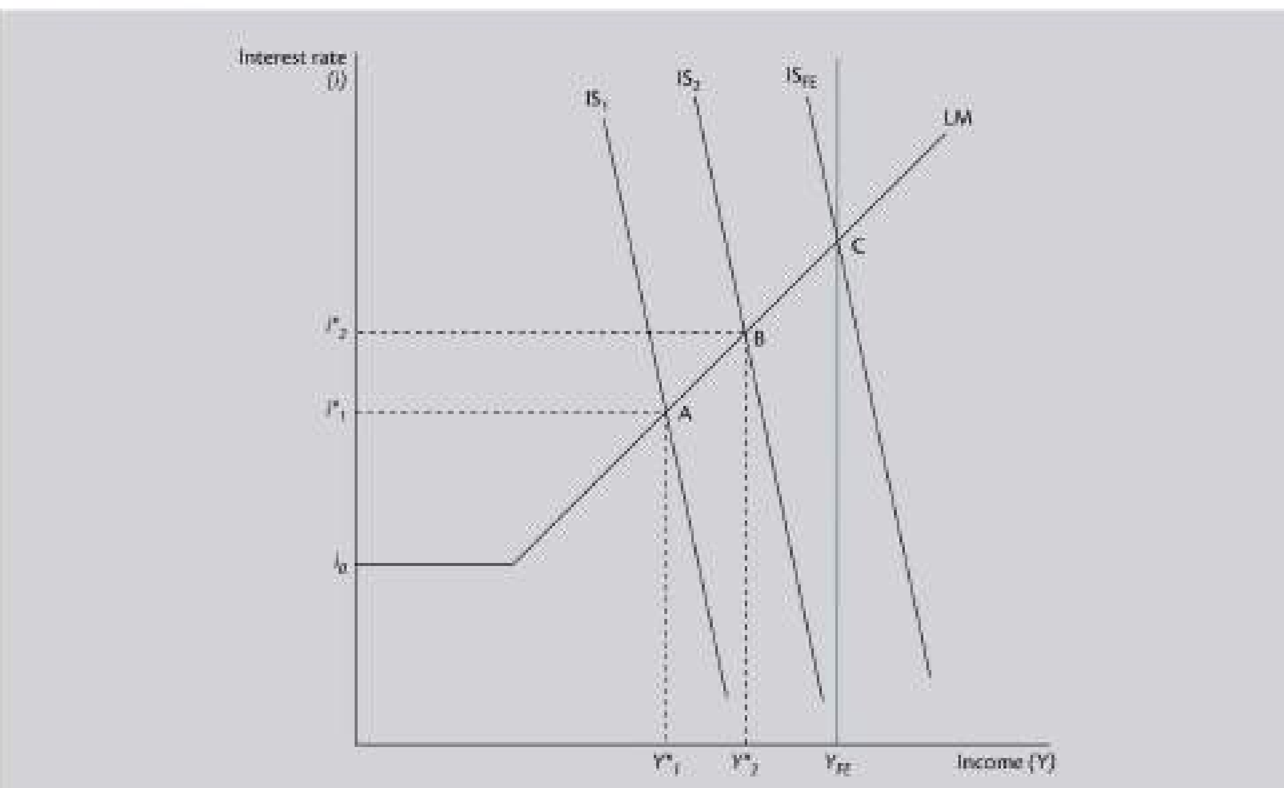
Similarly, a fall in government spending shifts the IS curve to the left because for a given interest rate, the equilibrium level of national income falls when autonomous spending falls.

The magnitude of the shift up or down in the IS curve resulting from a rise(fall) in autonomous spending is determined by the magnitude of the change in autonomous spending and the size of the expenditure multiplier. Thus, for a given change in autonomous spending, the shift in the IS curve will be larger, the larger is the value of the expenditure multiplier. Note that fiscal policy is still effective in a liquidity trap. The IS curve shifts to the right when government spending rises and shifts to the left when the government spending falls.

Figure 28.9 depicts expansionary fiscal policy, in the form of an increase in government spending. The initial equilibrium combination of the interest rate and national income is (i^*_1, Y^*_1). Point A shows the equilibrium position where IS_1 cuts the LM curve.

The treasury decides that the output gap (measured by the difference between the full employment national income level (Y_{FE}) and the current national income level (Y^*_1)) is too high (given the implied mass unemployment) and government spending is increased in order to stimulate aggregate demand.

Figure 28.9 Expansionary fiscal policy



The IS curve shifts to IS_2 which creates a new equilibrium at point B, with the equilibrium combination of the interest rate and national income at (i^*_2, Y^*_2). You will note that both the interest rate and national income are higher. The increase in national income arises because at higher levels of aggregate demand, firms produce more output (and hire more workers).

How do we explain the higher interest rates? Within the IS-LM framework, the rising national income that follows the increased aggregate demand raises the transactions demand for money. With the money supply fixed, the rising demand for money creates an excess demand for money at the original equilibrium interest rate (i^*_1) and rising interest rates motivate people to hold less cash. This is because the opportunity cost of holding wealth in the form of cash rises when interest rates rise.

You will note that if the interest rates had not increased, the expansion in income would have been greater than the shift from Y^*_1 to Y^*_2 and would be given by the simple multiplier, α, as outlined in Chapter 15.

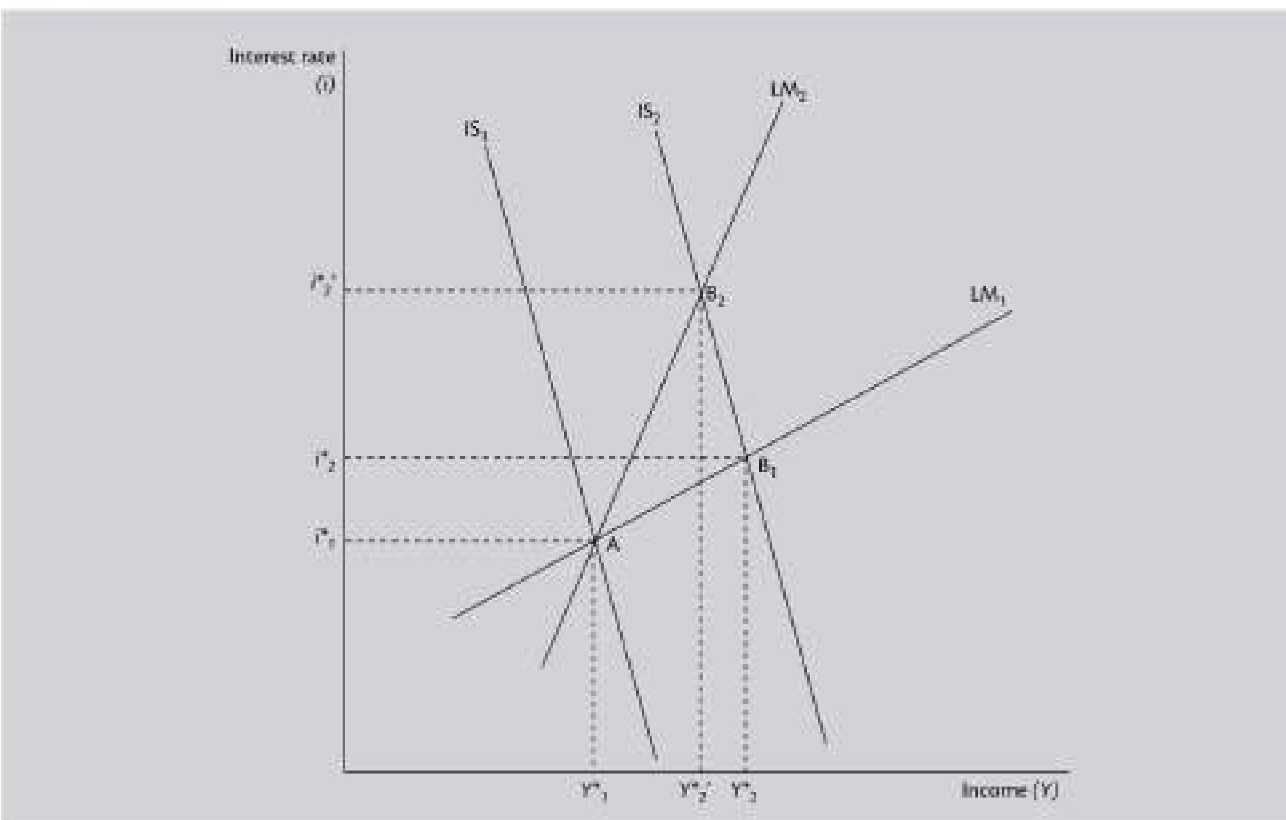
How do we explain the smaller multiplier effect in the IS-LM model? The rising interest rates impact negatively on private investment (and other interest sensitive components of aggregate demand) which offsets some of the increase in government spending. In the policy debates, this impact is referred to as **financial crowding out**. Crowding out can be measured as the amount by which investment (and other interest sensitive spending) falls when government spending increases.

Note that fiscal policy could achieve full employment (at point C) if the government kept increasing government spending such that the IS curve shifted to IS_{FE}.

The extent of the financial crowding out depends on the slope of the LM curve. The steeper is the LM curve, the less expansionary will be fiscal policy, and the larger is the crowding out effect.

This is demonstrated in Figure 28.10. From an initial equilibrium at point A, a fiscal stimulus which shifts the IS curve from IS_1 to IS_2 would increase national income to Y^*_2 and the interest rate to i^*_2 (equilibrium point B_1) with the flatter LM curve (LM_1).

Figure 28.10 Fiscal policy and financial crowding out



With a steeper LM curve (LM_2), for the same fiscal stimulus, the new equilibrium point is B_2 with a lower equilibrium income outcome ($Y^*_2{}'$) and a higher equilibrium interest rate ($i^*_2{}'$) than occurred at Point B_1. Thus, the extent of financial crowding out is higher with LM_2 than with LM_1.

What explains this difference? The LM curve is steeper the more sensitive the demand for money (transactions and precautionary motives) is to national income changes and the less sensitive the speculative demand for money is to changes in interest rates.

Thus, small changes in national income lead to large changes in excess money demand at a given money supply level and the rise in interest rates to restore money market equilibrium, other things being equal, has to be larger as a consequence.

An extreme position is complete financial crowding out. This would occur if the LM curve was vertical. In this situation, a given rise in government spending, for example, would be exactly offset by a decline in investment as the interest rate rose. In that situation, fiscal policy would be totally **ineffective**.

We should note that financial crowding out is not exclusively confined to fiscal policy changes. Any of the autonomous spending components that can shift the IS curve and increase national income can trigger the money market mechanisms that see interest rates rise and interest sensitive components of aggregate demand stifled.

Note that the other extreme position would be a horizontal LM curve at some given interest rate. In this case there would be no financial crowding out and the fiscal stimulus to aggregate demand would be fully translated into changes in national income. We will return to this extreme position when we discuss endogenous money theories later in the chapter.

If investment (and other components of aggregate demand) is not sensitive to interest rates, crowding out will be less. This is because even if the LM curve is upward sloping, so that government spending pushes up interest rates, investment is not reduced by much because it is not sensitive to interest rates. At the extreme, if aggregate demand were perfectly interest inelastic, there would be no financial crowding out.

To conclude this section:

- An expansionary fiscal policy increases national income and interest rates.
- Higher interest rates crowd out private investment and reduce the size of the multiplier.
- The extent of the financial crowding out depends on the slope of the LM curve. The steeper the LM curve is, the larger the crowding out effects will be, and hence fiscal policy will be less expansionary.
- There is complete crowding out when the LM curve is vertical and zero crowding out when the LM curve is horizontal.
- The steeper the IS curve, the smaller is the crowding out effect.

28.7 Introducing the Price Level: The Keynes and Pigou Effects

Our derivation of the IS-LM framework initially assumed that the price level is fixed and all changes in output are real. This is consistent with the simple income-expenditure model developed in Chapter 15 where the focus was on the manner in which output and employment respond to changes in aggregate demand.

We assumed that firms are willing to supply whatever is demanded up to full capacity without changing their prices. In this vein, we also treated the nominal and real interest rate as being interchangeable. In this section we consider how changes in the price level impact on output and interest rates.

The price level is introduced into the IS-LM framework as an exogenous variable, that is, determined outside of the interest rate income equilibrium defined by the intersection of the IS and LM curves. There are several complications involved in adopting this assumption which we will abstract from for the sake of simplicity.

The income-expenditure model developed in Chapter 15, which underpins the derivation of the IS curve was defined in real terms. Thus, the expenditure components (consumption, investment, government spending and net exports) are all measured in constant prices.

We would expect the IS curve therefore to be invariant to changes in the general price level given that households, firms, government and the external sector have made decisions regarding real expenditures.

BOX 28.1 JOHN HICKS ON HIS IS-LM FRAMEWORK

John Richard Hicks was a British economist who 'invented' the IS-LM general equilibrium macroeconomic framework.

In his 1937 article published in *Econometrica*, "Mr. Keynes and the 'Classics'; A Suggested Interpretation", Hicks sought to provide an interpretation of Keynes' General Theory within a single diagram, the IS-LM model. As the model became popularised and appeared in standard macroeconomic textbooks, the terminology became IS-LM to describe the product market equilibrium (IS) and the money market equilibrium (LM).

The IS-LM model was designed to demonstrate how the determination of total real output was dependent on a general equilibrium in the product and money markets.

Hicks said he "invented a little apparatus" (Hicks 1937: 157) (the IS-LM framework) to bring together Keynesian and Classical economics into an integrated model.

By the 1970s, Hicks started to sign his academic papers John Hicks rather than J.R. Hicks, which reflected his growing dissatisfaction with his earlier work.

In 1975, to formalise his transition away from his earlier views, he wrote: "J.R. Hicks ... [is] a 'neoclassical' economist now deceased ... John Hicks ... [is] a non-neo-classic who is quite disrespectful towards his 'uncle'" (Hicks: 1975: 365).

The issue was that he had begun to realise that the static equilibrium IS-LM model left out the key contribution of Keynes; the importance of time and endemic uncertainty.

For example, in the IS-LM model the current flow of investment is meant to be sensitive to interest rate changes in the same period, which is one way in which the money market outcome influences the product market equilibrium. But investment in any period is largely predetermined by decisions made in previous periods.

In 1980, Hicks wrote that he rejected the way in which his little apparatus had been deployed by economists and the policy interpretations that they had drawn from it.

He said:

> *The IS-LM diagram, which is widely, but not universally accepted as a convenient synopsis of Keynesian theory, is a thing for which I cannot deny that I have some responsibility. It first saw the light in a paper of my own, "Mr. Keynes and the Classics" (1937) ... I have, however, not concealed that, as time has gone on, I have myself become dissatisfied with it ... [the] diagram is now much less popular with me than I think it still is with many other people ... (Hicks 1980:139)*

Hicks had also come to realise that the IS-LM model contains an important logical flaw. The IS curve concerns flows of spending and income, while the LM curve concerns the demand and supply of money, which is a stock variable. The equilibrium point is a combination of income and interest rate that simultaneously satisfies equilibrium in both the goods and money markets, that is, simultaneously satisfies equilibrium of flows and stocks. In 1975 Duncan Foley demonstrated that a simultaneous equilibrium of stocks and flows is conceptually very complicated. To put it simply, a stock is at a point in time while a flow occurs over time. We can define a stock equilibrium for purposes of the LM curve on, say, 1 January of the year; but the flow equilibrium shown in the IS curve takes place over time, say from 1 January to 31 December. If that is the case, what is the time period that is represented in any of the IS-LM figures presented earlier in this chapter?

A better way to approach such questions is through what is called 'stock flow consistent modelling', as developed by Wynne Godley. But these are difficult issues, beyond the scope of this textbook. Suffice it to say here that the IS-LM approach is definitely not 'stock flow consistent', which is something Hicks began to recognise. Consequently, he rejected the use of the logically incoherent IS-LM model for policy analysis, which is what orthodox economists have primarily used it for!

(See also the quote from Hicks at the end of this chapter.)

However, to this point our analysis of the money market has finessed the question of the price level. In the orthodox model, the demand for money is a demand for real balances, motivated by the need to make transactions for the exchange of goods and services, which we have just noted are defined in real terms.

But, the money supply is specified in nominal terms, an amount of dollars, and forms the unit in which all the other variables are measured.

The real value of a given stock of money on issue, however, varies with the price level. For a given stock of dollars on issue, the real value is higher when the price level is lower, and vice versa.

For example, assume that the money supply on issue is $1,000 billion and the price index is one. The real value of the money supply would be $1,000 billion. Now if the price level rose by five per cent the price index would be 1.05 and the real value of the money supply would drop to $952.4 billion. This means that users of the currency have less available in real terms to use for purchases and speculative holdings.

The same contraction in the real value of the money supply could arise if the price level was unchanged (that is, the index remained at one) and the nominal money supply fell to $952.4 billion.

In other words, the real value of the money supply can fall if the price level rises (for a given nominal money stock) or if the nominal money stock falls (for a given price level).

Alternatively, the real value of the money supply can rise if the price level falls (for a given nominal money stock) or if the nominal money stock rises (for a given price level).

Within the logic of the IS-LM framework, it is clear that if the price level rises and reduces the real value of the money supply, the interest rate will rise because at the previous equilibrium interest rate, there will now be a shortage of real balances relative to the demand for them.

The introduction of the general price level modifies our LM curve derivation. If a rise in the price level (with a constant nominal money stock) is equivalent in real terms to a decline in the nominal stock of money (at constant prices) then we can capture this impact via shifts in the LM curve. The LM curve shifts to the left when the price level is higher, other things being equal, and to the right when the price level is lower.

Figure 28.11 depicts a family of LM curves with each individual curve corresponding to a different price level, with P_0 being the highest price level and P_3 the lowest. The introduction of the price level now means that the interest rate, income equilibrium is now contingent on the price level. This means that within this framework, the national income equilibrium can shift without any change in monetary or fiscal policy settings, if the price level changes.

This observation was central to the debates between Keynes and the Classical economists during the 1930s, which we examined in detail in Chapters 11 and 12.

Assume that the economy is currently at point A in Figure 28.11, where the interest rate is i_0 and national income is Y_0. The price level is P_0. The full employment output level is at Y_{FE}, so that the current equilibrium corresponds to what Keynes would refer to as an underemployment equilibrium. At point A, the product and money markets are in equilibrium but there is an output gap and there would be mass unemployment in the labour market as a consequence.

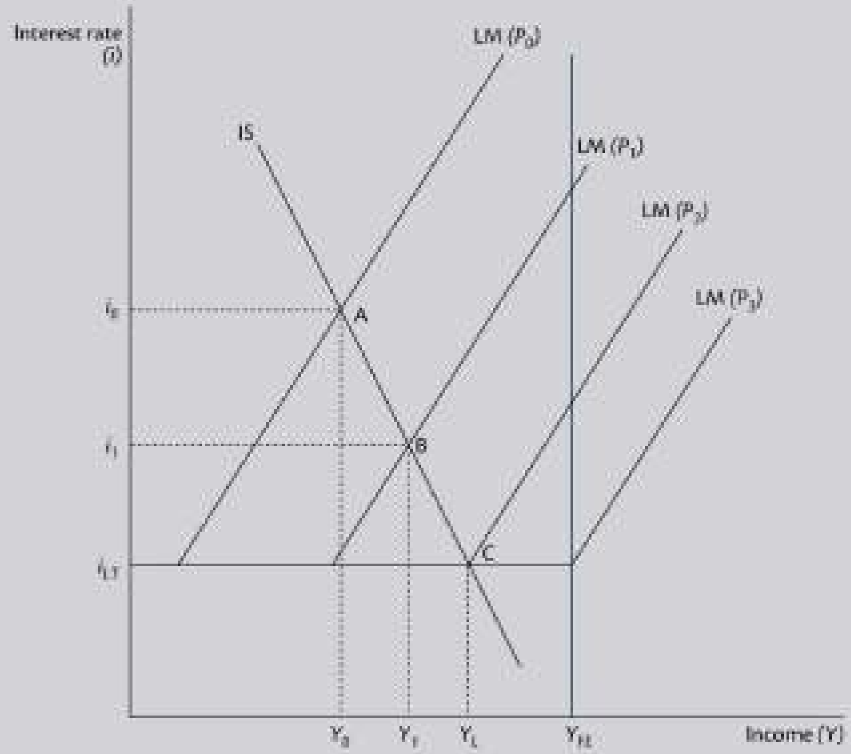
Keynes considered this to be the **general** case for a monetary economy and depicted the neoclassical model as a **special** case in which the equilibrium that emerged was consistent with full employment. For Keynes, a monetary economy could be in equilibrium at any level of national income.

The neoclassical response to this was that unless we impose fixed wages on the model, the persistent mass unemployment would eventually lead to falling nominal wages and prices.

While this might not lead to a fall in the real wage (if nominal wages and prices fall proportionately), which would negate the traditional neoclassical route to full employment via marginal productivity theory, the fact remains that the lower price level increases real balances in the economy.

The reasoning that follows is that the reduction in prices leads to a decline in the nominal transaction demand for money at every level of income, because goods and services are now cheaper. With the nominal stock of money fixed, the expansion of real balances combined with a decline in the demand for liquidity results in a decline in the rate of interest. As long as future expectations of returns are not affected adversely by the deflationary environment, the reduction in the rate of interest stimulates investment spending, which leads to increased

Figure 28.11 The Keynes effect



aggregate output and income via the multiplier effect. As long as there is an output gap, deflation will continue and the interest rate will continue to fall until the economy is at full employment.

The link between real balances and the interest rate is referred to as the **Keynes effect**.

In terms of Figure 28.11, the LM curve shifts outwards as the price level falls and the rising investment is depicted as a movement along the IS curve.

For example, if the price level fell to P_1, the LM curve would shift and a new IS-LM equilibrium would result at point B, with the interest rate at i_1 and national income at Y_1. Under the circumstances depicted, this is not a full employment level of national income.

As a result of this observation, the neoclassical economists argued that an underemployment equilibrium was a special case when wages and prices were fixed, given that flexible prices could reduce the output gap and unemployment via LM curve shifts.

The view that Keynes' underemployment equilibrium was a special case of the more general flexible price model became known as the neoclassical synthesis. This approach recognised that aggregate demand drives income and employment (the so-called Keynesian contribution), but that the economy would tend to full employment if wages and prices were flexible (the Classical contribution).

Note that the capacity of the Keynes effect to deliver output and employment gains is limited. If there is a liquidity trap (i_{LT}) then the maximum expansion in national income that is possible via falling prices would be Y_L at point C (where the IS curve intersects with the flat segment of the LM curve). At that point, there would still be unemployment and if wages and prices were flexible and behaved according to the Classical labour market dynamics, the price level would continue to fall, say to P_3. The LM curve would continue to shift out, but there would be no further expansion in national income beyond Y_L because the increase in real balances would not reduce the interest rate below i_{LT}.

The Classical route to full employment thus would require the full employment level of national income to lie at a point where the intersection of the IS-LM curves produced an equilibrium interest rate that was equal to or above i_{LT}.

The Keynes effect is so named because the expansion that follows a reduction in the price level occurs through a rise in aggregate demand, first through the interest rate stimulus to investment, and second through the standard expenditure multiplier inducing higher consumption expenditure.

However, as we learned in Chapter 12, Keynes did not support wage and price cuts as a way to achieve full employment. He considered the social consequences of wage cuts to be unacceptable and instead advocated increasing the nominal money supply as the way to increase real balances. But the limits to expansion posed by the possible existence of a liquidity trap dissuaded Keynes from considering the Keynes effect as a plausible route to full employment.

There are several other arguments that militate against a reliance on the Keynes effect for achieving full employment. Chapter 19 of Keynes' *General Theory*, which is devoted to the impacts of money wage changes on aggregate demand, presented several such arguments. Among other impacts, Keynes argued that lower money wages and prices will lead to a redistribution of real income, "(a) from wage earners to other factors entering into marginal prime cost whose remuneration has not been reduced, and (b) from entrepreneurs to rentiers to whom a certain income fixed in terms of money has been guaranteed" (Keynes 1936: 262). He concluded that the impact of "this redistribution on the propensity to consume for the community as a whole" would probably be more "adverse than favourable" (Keynes 1936: 262).

Moreover, if there are falling money wages and prices (Keynes 1936: 264): "the depressing influence on entrepreneurs of their greater burden of debt may partly offset any cheerful reactions from the reduction of wages. Indeed if the fall of wages and prices goes far, the embarrassment of those entrepreneurs who are heavily indebted may soon reach the point of insolvency — with severely adverse effects on investment." Overall, Keynes concluded that there was "no ground for the belief that a flexible wage policy is capable of maintaining a state of continuous full employment" (Keynes 1936: 267).

The debt deflation argument was also recognised by other economists such as Irving Fisher in 1933, Michal Kalecki in 1944 and Hyman Minsky in 1982.

The Classical view proposed an additional mechanism that could generate full employment as long as wages and prices are flexible. The so-called **Pigou effect** was named after Keynes' principal antagonist at Cambridge University, Arthur Pigou, whose work exemplified the Treasury View during the Great Depression. The Pigou effect is also known as the **real balance effect**.

While the Keynes effect works via interest rate responses to changing real money balances then stimulating investment, the Pigou effect is based on the view that falling prices would stimulate consumption expenditure (see Chapters 11 and 12). It was argued that the real value of household wealth rises as prices fall and this reduces the need to save. As a result the consumption function shifts upwards (higher levels of consumption at each income level) and this would shift the IS curve outwards.

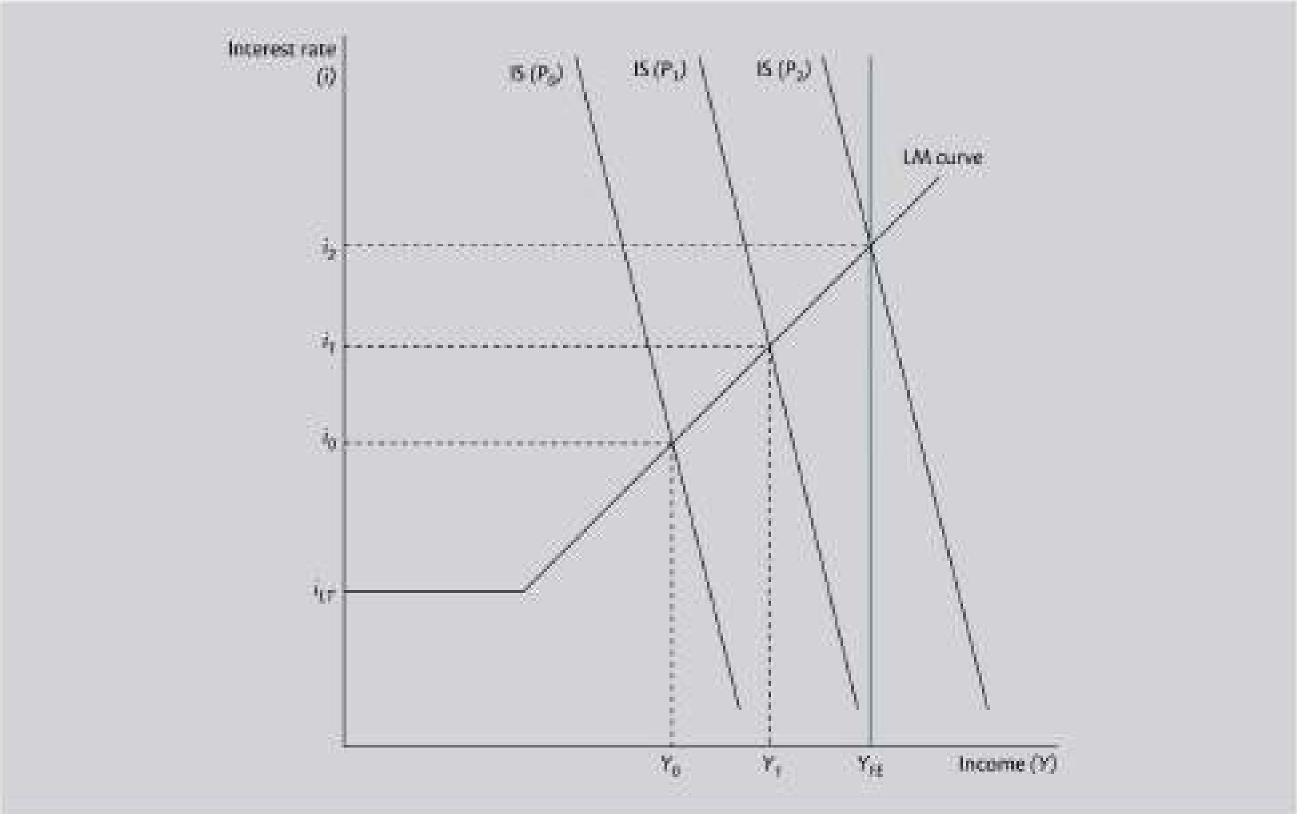
Figure 28.12 captures the Pigou effect. Note we abstract from any impacts on the LM curve of the falling price level to highlight the shifting IS curve. If we start from an initial underemployment equilibrium at i_0 and Y_0 with the price level at P_0. The argument is that wage and price levels would fall given the output gap ($Y_0 < Y_{FE}$) and this would increase real wealth balances and stimulate consumption, thus pushing the IS curve outwards and leading to an expansion in national income. Eventually, if prices were sufficiently downwardly flexible, the economy would achieve full employment at i_1 and Y_{FE}, with the lower price level, P_1.

You will note that unlike the Keynes effect, whose effectiveness is limited by the possibility of encountering a liquidity trap, the expansionary possibilities of the Pigou effect are unlimited.

The introduction of the Pigou effect provided a theoretical device to combat Keynes' argument that when aggregate demand is deficient (relative to the full employment level), wage and price flexibility would not guarantee full employment. However, studies have rejected its practical importance. The empirical research literature has found that wealth effects have tended to be small and insufficient to resolve a major recession.

Note also, that if wealth consists of claims on private sector households and firms, then falling prices benefit creditors, but hurt debtors. The real value of financial wealth is held by creditors, but that means it is harder for

Figure 28.12 The Pigou effect



debtors to service the debt out of falling nominal incomes. This is because debt contracts are normally written in nominal terms, not real terms. That would limit the effectiveness of a Pigou effect as the debtors cut their own spending to bear the burden of the rising real debt. It is even worse if they default on the debt, because in that case, even the creditors are hurt, and they might cut their spending too.

However, if wealth consists of claims on government (that is, if it takes the form of coins, central bank notes, and government bonds), then falling prices and rising real values of such debt will not (necessarily) cause the government to cut its own spending. In that case, the Pigou effect can still operate.

28.8 Limitations of the IS-LM Framework

There have been many critiques of the IS-LM framework over the years. Many have concentrated on whether the approach is a faithful representation of Keynes' General Theory, as was its initial purpose. Even its originator John Hicks accepted that it was not a valid depiction of Keynes' theories.

Other critiques have concentrated on issues relating to its static nature and the fact that it can tell us nothing about what happens when the economy is not in equilibrium.

A third objection relates to its denial of the realities of central bank operations and the way in which the commercial banks function.

In this section, we focus on the last two of these lines of attack.

The endogeneity of the money supply

First, IS-LM analysis relies on the assumption that the money supply is 'exogenous', that is, controlled by the central bank and thus, independent of the demand for funds.

The underlying theory supporting this assumption centres on the money multiplier, which we examined in detail in Chapter 10. The assumption is that the central bank is in control of the so-called **monetary base** (MB) (the sum of bank reserves and currency at issue), and the money multiplier transmits changes from the base into changes in the money supply (M_s). By setting the size of the monetary base, it is thus asserted that the central bank controls the money supply, as is depicted in the derivation of the LM curve.

As we learnt in Chapter 23, this conceptualisation of monetary operations is not remotely applicable to the real world. A senior official in the US Federal Reserve Bank of Boston, A.D. Holmes identified what he called "operational problems in stabilising the money supply" as far back as 1969:

> *The idea of a regular injection of reserves ... suffers from a naive assumption that the banking system only expands loans after the System (or market factors) have put reserves in the banking system. In the real world, banks extend credit, creating deposits in the process, and look for the reserves later. The question then becomes one of whether and how the Federal Reserve will accommodate the demand for reserves. In the very short run, the Federal Reserve has little or no choice about accommodating that demand; over time, its influence can obviously be felt (Holmes 1969: 73).*

The reality, which we analysed in detail in Chapter 23, is that the central bank sets the so-called official policy or target interest rate. This is the rate at which it is prepared to provide funds to the banking system on an overnight basis.

We summarise the key points below:

- Bank loans create deposits, that is, banks react to the demand for credit from borrowers rather than on lending deposits.
- The demand for credit depends on the state of economic activity and the level of confidence in the future.
- Bank lending is not constrained by reserve holdings. The reserves are added on demand by the central bank where needed.
- Rather than driving the money supply, the monetary base responds to the expansion of credit by the banks.
- This process means the money supply is endogenously determined and the central bank has no real capacity to maintain any quantity targets.

The fact that the money supply is endogenously determined means that the LM schedule will be horizontal at the policy interest rate. All shifts in the interest rates are thus set by the central bank and funds are supplied elastically at that rate in response to the demand. In this case, shifts in the IS curve would not impact on interest rates. From a policy perspective this means the simple notion that the central bank can solve unemployment by increasing the money supply is flawed.

If the central bank tries to increase reserves in a discretionary manner, this would only result in excess reserve holdings and push the overnight interest rate toward zero without actually increasing the money supply. To avoid this, the central bank would have to pay the policy rate on those excess reserves.

Unemployment is typically the result of high liquidity preference, people want to hold cash rather than spend it, given uncertainty about the future. In those cases the demand for loans collapses and the banks become more cautious as to whom they will loan funds for fear of losses. Under these conditions, the central bank cannot simply increase the supply of money to raise aggregate demand.

Since the GFC, central banks have added massive volumes of reserves to the banking system via the so-called quantitative easing programmes, which we analysed in detail in Chapter 23. The demand for funds was so subdued that credit expansion also slowed dramatically and the banks were content to hold vast quantities of low interest bearing reserves.

Note that most orthodox monetary economists have given up the notion that the central bank can control the money supply, accepting that policy sets the overnight rate target. They have dropped the LM curve from their models and developed a New Monetary Consensus model to replace the IS-LM model. We discuss this in Chapter 30.

Expectations and time

Consider the role of the investment function in the derivation of the IS curve. Investment is said to be dependent on the interest rate (cost of funds) and perhaps output (via the accelerator affect, see Chapter 25).

While the IS-LM approach of John Hicks tried to represent what he saw as the key elements of Keynes' General Theory, it clearly left out issues relating to uncertainty and probability that Keynes saw as being crucial in the way that long-term expectations were formed. Chapter 12 of the *General Theory* was devoted to this topic. In it, Keynes wrote:

> *The outstanding fact is the extreme precariousness of the basis of knowledge on which our estimates of prospective yield have to be made. Our knowledge of the factors which will govern the yield of an investment some years hence is usually very slight and often negligible. If we speak frankly, we have to admit that our basis of knowledge for estimating the yield ten years hence of a railway, a copper mine, a textile factory, the goodwill of a patent medicine, an Atlantic liner, a building in the City of London amounts to little and sometimes to nothing; or even five years hence (1936: 149–50).*

Thus, the decision to invest is dependent on the "state of long-term expectation" (Keynes 1936: 147), which is ignored in the static IS-LM approach.

Investment, among other key economic decisions, is a forward-looking process, where firms form guesses about what the state of aggregate demand will be in the years to come. It is necessary because the process of creating new capital stock is lengthy, and involves a number of separate decisions, type of output to produce, nature of capital required to produce it, design, access to supply and ordering, and quantum, are all separated in time.

The investment spending today is the result of decisions taken in some past periods about what the state of the world will be today and into the future. Investment spending is not a tap that is turned on or off when current interest rates change.

The psychological factors that are crucial for comprehending the decision to consume (marginal propensity to consume); the decision to invest (marginal efficiency of capital); and the determination of the labour market wage bargain (implicit in the IS-LM approach) are abstracted from in the derivation of the equilibrium. Thus, dynamic processes with complex feedback loops are frozen in time by the need to derive static IS and LM curves.

The failure to include the crucial role of expectations and historical time means that the IS-LM framework is reduced to presenting a general equilibrium static solution that has little place in a dynamic system where uncertainty is a key driver in economic decision making.

The last word in this chapter will go to the original architect of the IS-LM approach, John Hicks, who reflected on his creation and the way it had been subsequently used in a 1980 article in the *Journal of Post Keynesian Economics*:

> *I accordingly conclude that the only way in which IS-LM analysis usefully survives – as anything more than a classroom gadget, to be superseded, later on, by something better – is in application to a particular kind of causal analysis, where the use of equilibrium methods, even a drastic use of equilibrium methods, is not inappropriate. I have deliberately interpreted the equilibrium concept, to be used in such analysis, in a very stringent manner (some would say a pedantic manner) not because I want to tell the applied economist, who uses such methods, that he is in fact committing himself to anything which must appear to him to be so ridiculous, but because I want to ask him to try to assure himself that the divergences between reality and the theoretical model, which he is using to explain it, are no more than divergences which he is entitled to overlook. I am quite prepared to believe that there are cases where he is entitled to overlook them. But the issue is one which needs to be faced in each case.*
>
> *When one turns to questions of policy, looking toward the future instead of the past, the use of equilibrium methods is still more suspect. For one cannot prescribe policy without considering at least the possibility that policy may be changed. There can be no change of policy if everything is to go on as expected – if the economy is to remain in what (however approximately) may be regarded as its existing equilibrium. It may be hoped that, after the change in policy, the economy will somehow, at some time in the future, settle into what may be regarded, in the same sense, as a new equilibrium; but there must necessarily be a stage before that equilibrium is reached. There must always be a problem of traverse. For the study of a traverse, one has to have recourse to sequential methods of one kind or another. (Hicks 1980: 152–3, emphasis in original)*

The last point was telling. While the intersection of given IS and LM curves might reflect conditions now, the other points on the respective curves are what John Hicks called "theoretical constructions" (1980: 149) and "surely do not represent, make no claim to represent, what actually happened" (1980:149).

Conclusion

This chapter has examined in detail the Neoclassical Keynesian approach that dominated macroeconomics for over a quarter of a century after the Second World War. While supposedly a "Keynesian" approach, in reality it was largely based on orthodox economics. It used a simplistic IS-LM model that was based on simultaneous clearing in the money and goods markets. Deviations from a full employment equilibrium could occur, but market mechanisms would (eventually) move the economy to full employment. Not only did this view dominate in academia, but it also formed the basis of much policymaking.

Because of policy, predictive, and theoretical failures, the IS-LM model was gradually dropped by academics (although it still resides in the minds of many policymakers and a few influential professors of economics). In Chapter 30 we will discuss the New Macroeconomic Consensus model that has replaced IS-LM as the dominant model. We will see that while there are differences between the two, there remains a similar equilibrium framework.

References

Fisher, I. (1933) "The Debt-Deflation Theory of Great Depressions" *Econometrica*, 1(4), 337–57.

Foley, D.K. (1975) "On Two Specifications of Asset Equilibrium in Macroeconomic Models", *Journal of Political Economy*, 83(2), 303–24.

Hicks, J.R. (1937) "Mr. Keynes and the 'Classics'; A Suggested Interpretation", *Econometrica*, 5(2), 147–59.

Hicks, John (1975) "Revival of Political Economy: The Old and the New", *The Economic Record*, 51(135), 365–7.

Hicks, J. (1980) "IS-LM: 'An Explanation'", *Journal of Post Keynesian Economics*, 3(2), 139–54.

Holmes, A. (1969) "Operational Constraints on the Stabilization of Money Supply Growth. In Controlling Monetary Aggregates", *Federal Reserve Bank of Boston*, 65–77.

Kalecki, M. (1944) "Professor Pigou on the 'Classical Stationary State' A Comment", *The Economic Journal*, 54(213), 131–2.

Keynes, J.M. (1936) The General Theory of Employment, Interest, and Money, London: Macmillan, 1957 Reprint.

Minsky, H. (1982) "Debt Deflation Processes in Today's Institutional Environment", *Banca Nazionale del Lavoro Quarterly Review*, 143, 377–93.

Endnotes

1. In Figure 28.4, the upward sloping section of the LM curve is shown as being linear. It is upward sloping, but not necessarily linear.
2. Dennis Robertson, who worked closely with Keynes at Cambridge University in the 1930s, coined the term *liquidity trap*.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER 28 APPENDIX: THE IS-LM ALGEBRA

Simplified open economy

The formal IS-LM model for a simplified open economy begins with the following relationships:

(28.3)	$Y = C + I + G + X - M$	National income identity
(28.4)	$C = C_0 + cY_d$	Consumption function
(28.5)	$Y_d = Y - T$	Disposable income
(28.6)	$T = tY$	Tax revenue (proportional tax rate, t)
(28.7)	$I = I_0 - bi$	Investment function
(28.8)	$G = G$	Government spending
(28.9)	$M = mY$	Import function

Product market equilibrium

The product market equilibrium can be solved as a relationship between GDP (Y) and the interest rate (i) given the autonomous spending aggregates and the value of the multiplier. Substituting Equations (28.4) to (28.9) into (28.3) we get:

$$Y = C_0 + cY - ctY + I_0 - bi + G + X - mY \tag{28.10}$$

Rearranging gives the equation for the IS curve:

$$Y = \alpha(A - bi) \tag{28.11}$$

where $\alpha = 1/(1 - c(1 - t) + m)$, the expenditure multiplier and A is the autonomous spending component, $C_0 + I_0 + G + X$.

The slope of the IS curve is given by αb, so the larger the multiplier (α) and the sensitivity of investment to interest rates (b), the flatter will be the slope because the response of national income to a given interest rate will be larger.

Money market equilibrium

The money market equilibrium is given by the equality of money supply and money demand.

$$M_s = kY - hi \tag{28.12}$$

which yields the LM curve where Y is a function of i:

$$Y = (1/k)M_s + (h/k)i \tag{28.13}$$

The slope of the LM curve is given by (h/k), so the larger the sensitivity of the demand for money to interest rates (h) and the smaller the sensitivity to income (k), the flatter will be the slope because for a small change in interest rates, a much larger change in national income will be required to maintain the equality between the demand for money and the given money supply.

Equation (28.13) can also be written with i as a function of Y:

$$(28.14) \quad i = (k/h)Y - (1/h)M_s$$

General equilibrium

A state of general equilibrium in this context is defined as the interest rate and income level that generates simultaneous equilibrium in both the product and money markets. This equilibrium position corresponds to the intersection of the IS and LM curves.

To solve for the equilibrium level of national income we can substitute Equation (28.14) into the IS curve Equation (28.11) to give:

$$(28.15) \quad Y = \alpha[A - (b/h)(kY - M_s)]$$

Solving for equilibrium Y gives:

$$(28.16) \quad Y = \left(\frac{\propto}{1+\frac{\propto bk}{h}}\right)\left[A + \frac{b}{h}M_s\right]$$

Equation (28.16) indicates that equilibrium income is determined by autonomous spending (A), which includes the fiscal policy parameter (G), and the money supply (M_s).

We use the solution in Equation (28.16) to solve for the equilibrium interest rate:

$$(28.17) \quad i = \frac{\alpha}{h}\left(\frac{k}{1+\frac{\propto bk}{h}}\right)A - \left(\frac{1}{h+\propto bk}\right)M_s$$

Equation (28.17) tells us that the equilibrium interest rate is determined by autonomous spending (A) and the money supply (M_s).

Some economists use Equation (28.16) to define a fiscal policy multiplier, which indicates the change in national income for a given change in government spending if the money supply is held constant.

The fiscal policy multiplier is given by the coefficient on autonomous spending (A) in Equation (28.16). You will note that this is different to the simple expenditure multiplier (α) because it takes into account the impact of rising income on the interest rate and hence the investment spending that results from the shifts in the demand for money.

The simple expenditure multiplier is derived on the assumption that the interest rate does not impact on investment (that is, b can be treated as zero).

Similarly, it is possible to derive a monetary policy multiplier which shows the increase in national income for a given change in the money supply if government spending and tax rates are held constant.

This is given by the coefficient on M_s in Equation (28.16). Note though that this assumes that monetary policy is conducted through the central bank exercising its assumed control over the money supply. This is one of the flaws of the IS-LM framework when applied to the real world, the central bank does not have control over the money supply and the assumed money multiplier does not exist in any form other than as an *ex post*, non-causal accounting statement.

CHAPTER

29 MODERN SCHOOLS OF ECONOMIC THOUGHT

Chapter Outline

Learning Objectives

- Understand the basic theoretical framework of each of the main modern schools of thought.
- Be able to explain the main differences between the orthodox and heterodox approaches to macroeconomics.

29.1 Introduction

In the post-Second World War period, so-called Keynesian economics was dominant in the developed world with policymakers active in their use of both monetary and fiscal policy. These countries experienced high rates of employment and low inflation over a sustained period.

The oil price shocks of the 1970s and the tightening of macroeconomic policy led to the emergence of stagflation and provided an ideal platform for the Monetarist ideas of Friedman, which built on (neo)classical theory to gain credibility, as was outlined in Chapter 18.

This period in the 1970s proved to be a watershed with respect to macroeconomic theory and policy with the restoration of the dominance of free market principles, as articulated in New Classical theory. This model built on Monetarism, and latterly the New Monetary (Macro) Consensus which meant that active macroeconomic policy was confined to monetary policy. In Chapter 30, we provide a more detailed analysis of the New Monetary Consensus, which can be seen as a synthesis of many of the core ideas developed in this chapter.

On the other hand, while New Keynesian Theory accepted that flexible wages and prices led to market clearing and hence full employment, it was premised on the simple point that there were valid arguments as to why wages and prices were not flexible, so that there was a greater role for discretionary policy. The dominance of these free market ideas was confounded by the events leading up to, and the persistence of, the Global Financial Crisis (GFC), which we outline in Chapter 27.

In this chapter, we briefly summarise these different schools of thought. We discuss modern heterodox economics in terms of their key differences from these schools of thought.

29.2 The Rise of New Classical Economics

Roots in Friedman's Monetarism

In spite of its major flaws, the 'Keynesian' Neoclassical Synthesis which was outlined in Chapter 28 did allow for the possibility of an unemployment equilibrium resulting from insufficient aggregate demand.

In the aftermath of the Great Depression, post-war economists recognised a positive role for government's use of fiscal policy to close 'demand gaps' by increasing spending (or tax reduction). While monetary policy could also be used, the Neoclassical Synthesis economists doubted its efficacy, especially in a recessionary environment.

Monetary policy would work only if it could encourage firms and households to borrow and spend, which would be unlikely if their expectations were pessimistic. For this reason, most of the responsibility for raising aggregate demand was placed on fiscal policy.

Beginning in the late 1950s, Milton Friedman (often working with Anna Schwartz) challenged this belief with the development of an alternative approach, called Monetarism. He argued that monetary policy was actually far more potent than most economists believed at the time. He claimed that empirical evidence showed that changes in the money supply caused movements of nominal GDP. Also, he asserted that the 'invisible hand' of the market was more powerful than the Keynesians recognised.

Friedman reformulated the pre-Keynesian Quantity Theory of Money to make it more consistent with the approach used by the Neoclassical Synthesis Keynesians in deriving the LM curve. In his version, an increase of the money supply by the central bank causes spending and prices to rise sufficiently to raise money demand until equilibrium is restored between money demand and money supply.

With flexible wages and prices, the economy would tend towards full employment, so most of the adjustment to an increase in the money supply would be through rising prices, not real output. Hence, while accepting a more 'Keynesian' version of money demand, Friedman revived the Quantity Theory of Money and integrated it with a more thoroughly neoclassical 'free market' model than the Neoclassical Synthesis 'Keynesian' approach.

Friedman (1970) is perhaps best known for his claim that:

> *[I]nflation is always and everywhere a monetary phenomenon ... A steady rate of monetary growth at a moderate level can provide a framework under which a country can have little inflation and much growth. It will not produce perfect stability; it will not produce heaven on earth; but it can make an important contribution to a stable economic society. (p. 11)*

This statement reflects his belief that an increase of the money supply is a necessary and sufficient condition for prices to rise.

Friedman claimed that, in the long run, money is 'neutral', so that it only affects only nominal values of wages, prices, and interest rates, but in the short run it can have effects on the real economy.

In his 1968 paper, Friedman criticised the 'Keynesian' Phillips curve trade-off between inflation and unemployment, arguing that this can only apply in the short run. If the monetary authorities increase the rate of growth of the money supply, this raises the rate of increase of nominal wages and prices, which should not change behaviour.

We have given a lengthy exposition of Friedman's 'expectations augmented' Phillips curve (EAPC) which is also referred to as the 'accelerationist hypothesis' in Chapter 18, so to maintain continuity in the development of New Classical Economics, we shall just provide a set of summary points here.

- Under the EAPC, there is a short-run trade-off between inflation and unemployment because workers' expectations of price and wage movements are adaptive in the light of past price and wage outcomes, so workers can be temporarily fooled with respect to interpreting price and wage movements. Thus, changes in the money supply can have real effects on the economy.

- The natural rate of unemployment, that is, the level of unemployment that arises due to frictions in the labour and other markets, is associated with stable inflation.
- An unemployment rate that is less than the natural rate causes an accelerating (rising) rate of inflation. Thus, the use of discretionary fiscal (or monetary) policy to reduce unemployment below its natural level is counterproductive.
- In the absence of fooling, the Phillips curve would be vertical, with no trade-off between unemployment and inflation because raising the growth rate of the money supply would simply cause inflation to rise.

Some neoclassical economists wondered if this type of expectations formation was consistent with rational behaviour. First, both money supply data and inflation estimates are frequently reported. Second, 'rational' people should try to predict the central bank's policy moves in advance, so that expectations ought to be 'forward looking'. In that case monetary policy would have even less impact on real variables such as unemployment.

In any event, Friedman did not recommend the use of monetary policy in a discretionary manner. He proposed that the authorities simply target a low, but stable growth rate for the money supply; say, three per cent per year. If the velocity of money trends upward at a fairly constant rate of say, one per cent per year, and if real output grows on trend at four per cent per year, then a money supply growth rate of three per cent per year would mean a zero rate of inflation on average. We can write this as:

$$(29.1) \quad g_M + g_V = g_P + g_Y$$

where g stands for growth rate, M is money supply, V is velocity, P is aggregate price level, and Y is real (inflation adjusted) national income or GDP. A constant but higher growth rate of the money supply would generate a constant but positive rate of inflation because Equation (29.1) can be rewritten as:

$$(29.2) \quad g_P = g_M - (g_Y - g_V)$$

A stable inflation rate would not affect behaviour because no one would be fooled once inflationary expectations had adjusted to the positive rate of inflation.

Unlike Neoclassical Synthesis 'Keynesians', Friedman doubted that fiscal policy has much impact on GDP. The main reason is that an increase of government spending (or reduction of taxes) is likely to 'crowd out' private spending, through rising interest rates, as was explained in Chapter 28.

We can take this one step further if we assume a particular kind of rationality. Suppose that consumers believe that a government deficit this year (caused by either more government spending or by a tax cut) means higher taxes next year to 'pay off' any government debt issued. This is the idea behind 'Ricardian Equivalence'.

The 'rational' thing to do then is for households and firms to cut spending today in order to save for the future tax bills. Combining all these arguments, one concludes that fiscal policy is impotent. It would not raise output or employment because any increased government spending would lead to less private spending.

To conclude:

- Friedman argued that fiscal policy has little net effect on aggregate demand.
- Monetary policy does affect aggregate demand, but most of its impact is on the price level.
- Real output and employment are affected by monetary policy only due to temporary 'fooling'.
- The best form of monetary policy is to simply target a constant rate of growth of the money supply, which eliminates fooling.
- According to Monetarists, the invisible hand is quite effective at keeping the economy close to full employment growth, so macroeconomic policy can play a minor role.

New Classical Economics

By the early 1970s a more extreme anti-Keynesian approach than that taken by Friedman had emerged. Indeed, it was more fundamentally neoclassical than the old neoclassical approach that Keynes had criticised. It was referred to as the New Classical approach, and was based on a number of propositions:

1. Markets clear instantaneously. Thus, rather than moving slowly to equilibrium, all markets are always in equilibrium as prices adjust so that the quantity demanded equals the quantity supplied.
2. Agents, namely households and firms, have *rational expectations*. This is defined in a very specific way: the subjective expectations of all economic variables will coincide with the true or objective mathematical conditional expectations of those variables. This means that, while the agents do not have perfect foresight, any errors that they make are random. On average, expectations are correct.
3. Friedman's adaptive expectations are backward looking so that agents make systematic errors as their expectations only gradually come into line with reality. On the other hand, the assumption of rational expectations ensures that agents do not form systematically incorrect expectations.
4. The business cycle is due to 'surprises' that affect aggregate supply in a random fashion. Economic policy is effective only if it cannot be predicted. This is called the (Robert) Lucas supply function, according to which 'shocks' to aggregate demand are induced by random policy changes.
5. This last point requires some explanation. Friedman argued that monetary policymakers can 'fool' people into working and producing more real output because they mistake nominal wage and price increases for a real (relative) wage and price increase. Since expectations are formulated adaptively, it takes some time for them to realise they have been fooled. However, such behaviour is not rational, because as long as policy is predictable, it will be expected.
6. Only random policy would be unpredictable and thus capable of fooling anyone. (Remember that under *rational expectations*, individuals can make wrong predictions, but on average, expectations are correct because the errors are random.)

Deviations away from the natural full employment rate of unemployment are thus due to mistakes, or specifically to monetary policy surprises. Anticipated monetary policy changes will have no real effects; only unanticipated money matters in the sense that it affects decisions and thus real output.

Money is neutral in both the long run and the short run, with the exception that random monetary policy can have a real effect in the short run.

The pursuit of random policy however, would serve no purpose. To be effective, policymakers would have to be just as likely to increase the money supply as they are to decrease it in any given circumstance. Lucas (1972) concluded that Friedman's constant rate of growth of money supply rule made the most sense for policymakers. There is no point in trying to use monetary policy to influence the economy since predictable policy has no impact.

What about fiscal policy? The New Classical approach adopts the Ricardian Equivalence argument. Thus, a policy of fiscal stimulus would be impotent because any increase of government spending or tax cuts would be matched by an increase of current private sector saving in order to pay the anticipated future taxes.

In any case, fiscal stimulus is not needed if all markets clear continuously, which means that the economy is always at full capacity utilisation. There are no demand gaps that need to be filled.

This does not mean that government policy has no impact on the economy. New Classical economics, like neoclassical theory in general, is a supply side approach. If a policy can affect preferences, technology, or labour productivity, it can have real effects.

For example, improved education and training could raise labour productivity and thus real output; higher taxes might reduce the incentive to work and hence lower real output; the elimination of minimum wage laws would free the labour market to increase employment and real output. Since supply creates its own demand (Say's Law, see Chapter 12), there is no problem of effective demand.

Robert Lucas contributed another lasting argument: the Lucas Critique. In the 1960s it was fashionable to build large macro models of the economy, based on the Neoclassical Synthesis 'Keynesian' theory. These were used for short-term forecasts as well as for policy analysis.

The models did reasonably well in forecasting economic growth a quarter or two ahead, because they were driven by a stable consumption function. Of importance was the fact that the consumption equation in these models did not include an inflation term as one of the explanatory variables, even though theory might suggest that households would typically consider expected inflation to be important in their consumption decisions. The absence of an inflation variable was fine as long as inflation was low or stable as it was in the 1960s. As a result, the forecasting errors were small and the models proved to be useful to policymakers. In the more turbulent times since the early 1970s, as inflation began to rise after the OPEC oil price rises, the models started to generate large forecast errors until the omissions were taken into account.

However, the Lucas Critique concerned the use of models for policy analysis. These models use a combination of theory and econometric estimates to obtain a good fit of the data. They are necessarily backward looking as the parameters of the behavioural equations (such as the consumption function) are estimated using historical data.

From these estimates we can calculate, for example, a spending multiplier. To analyse the impact of an increase of government spending, we use the estimated size of the spending multiplier. Assume that an increase of an autonomous component of spending would be associated with an increase of national income based on an estimated multiplier of two. We then use that estimate to claim that a current increase of government spending by three billion dollars would generate growth of GDP by six billion dollars. Lucas argued that this is illegitimate because it presumes behaviour would not change in the face of the policy change. If people hold *rational expectations*, they will react to the prospective policy change.

For example, as discussed above, they might decide to consume less (save more) in anticipation of future higher taxes that they believe the government will eventually impose to pay down debt incurred now due to the higher government spending. To do policy analysis, the agents in the model must be presumed to hold rational expectations; those agents will have 'a model of the model' that they will use to formulate their expectations.

If our real world economy adheres to Ricardian Equivalence, then the model used by the agents must incorporate Ricardian Equivalence. More generally, the model builder must assume that agents hold *rational expectations* and a model that is consistent with the model used by the modeller! While in practice this proved to be difficult to achieve, economic modellers took this critique seriously and tried to introduce rational expectations into their models, but with little success.

While the assumption of *rational expectations* seems to be the critical innovation of the New Classical school, the more important assumption was that of continuous market clearing. As we will see, the New Keynesians were able to restore some Keynesian features to their theory, even though they adopted the seemingly radical assumption of *rational expectations*.

29.3 Real Business Cycle Theory

As discussed, both the Monetarist approach and the New Classical theory rely on a type of *fooling* by the central bank to explain fluctuations of national output and employment. While the New Classical approach is more 'neoclassical' than Friedman's Monetarism, which accepted some features of the Neoclassical Synthesis framework, it still relies on the non-neutrality of money in the short run. In this case, it is due to unanticipated policy.

Some neoclassical economists wanted to be even more true to rigorous neoclassical foundations, insisting that money is nothing but a medium of exchange that should not affect behaviour.

Instead, the business cycle must be explained by 'real' variables, such as changes to resource endowments, preferences, and productivity. This led to the development of Real Business Cycle theory (RBC). The most important assumptions of RBC are:

1. Rational expectations.
2. Continuous market clearing.
3. Monetary policy only determines nominal prices; money is always neutral.
4. There are large random fluctuations of preferences and technology.

Together these assumptions lead to several important conclusions. Fluctuations in employment are the consequence of individual utility maximisation behaviour. These outcomes are thus voluntary and optimal, given real wages and real interest rates: those who are not working have chosen not to work. There is no distinction between the short run and the long run with the economy always in equilibrium.

There are two factors driving changes to output and employment. The first is that the economy will grow on trend due to the growth of inputs to the production process, namely labour force growth, the accumulation of capital, and the rising exploitation of natural resources. We can assume that all these inputs grow at a reasonably stable rate, so the economy would grow at a reasonably constant rate if it were not for the second factor, namely random shocks to preferences and technology.

Random shocks cause the growth rate to deviate from trend. They can be positive or negative, shifting the economy to a higher or lower trend growth rate. Since these are random shocks, they are serially uncorrelated. Thus, a positive shock is just as likely to be followed by another positive shock as it is to be followed by a negative shock.

This leads to an entirely new way of looking at the business cycle. Until RBC theory, economists had always viewed the cycle as undesirable. The economy cycles between expansion and recession, and while expansions yield the benefits of higher employment and incomes, recessions have adverse consequences, particularly for the disadvantaged. Mainstream economists considered both the expansion and the recession phases as movements away from equilibrium.

Both Monetarist and New Classical approaches attributed the expansion phase to 'fooling' by the monetary authorities, who increased the rate of growth of the money supply beyond what was expected. Equilibrium would be restored at a lower real growth rate, but a higher inflation rate, once expectations were realigned with reality.

A recession would be induced by tighter monetary policy (lower money growth) than expected, and again equilibrium would be restored after expectations had adjusted. Heterodox economists such as Keynes and Veblen also addressed the cyclical nature of capitalist economies. Keynes developed his theory of fluctuations of effective demand and Veblen analysed the business enterprise, but they did not attribute these fluctuations solely to monetary policy.

Advanced treatment of the RBC model

Traditional business cycle theory can be specified as follows:

$$(29.3) \qquad Y_t = g_t + bY_{t-1} + Z_t$$

where g is the trend growth of output; b is a reversion parameter such that $0 < b < 1$; Z_t are random shocks; and t is the time subscript. This equation represents autoregression about a trend in the presence of a lagged output variable.

Fluctuations occur due to demand or supply shocks. These cause a temporary deviation from trend, with growth then reverting to the trend. Thus, a positive(negative) shock causes the growth path to shift upward(downward), signifying a business cycle expansion(contraction), which gradually reverts to the trend growth rate.

RBC advocates claim that a random walk with drift can better explain the cycle:

$$(29.4) \qquad Y_t = g_t + Y_{t-1} + Z_t$$

Note there is only one difference from the equation above, which is that b is set equal to one. As such, there is no reversion to trend; output has a unit root ($b = 1$). Here the random shock (Z_t) provides the 'random walk', while the trend growth term (g) generates the 'drift'.

Unlike the model in Equation (29.3), a productivity shock shifts the growth trend, where it remains until the next shock. These shocks are presumed to be random with no serial correlation. That means that it is impossible to predict whether a positive shock will be followed by a negative or a positive shock.

RBC's proposition is that if the economy moves as a random walk with drift, and if shocks to productivity growth are frequent and random, they might explain the apparent cycle.

RBC theorists argue that the cycle should not be represented as a deviation from trend, but a fluctuation of the natural rate of growth in response to a shock. Every point is optimal. Thus, they reject the notion of a business cycle altogether.

This does not mean that people are necessarily happy about choosing to leave their jobs after a negative shock, but that it is their utility-maximising response. Given the lower real wages that result from negative policy shocks, workers rationally choose more leisure over working.

A positive technology shock (the impulse) raises productivity. This shifts the production function upward, which increases the marginal productivity of labour. That shifts labour demand out, increasing employment, real output and the real wage. This is called propagation. The persistence is due to labour supply remaining higher until a negative productivity shock comes along and reduces the real wage.

Since market forces are believed to operate effectively to move the economy to equilibrium, RBC theorists assume continuous market clearing, albeit with variations of employment reflecting optimal choices between work and leisure.

Further, they argue that their theory fits the facts better than does the old, long-standing 'business cycle' approach. If the economy does cycle around a trend, one expects that a period when GDP growth is above trend (or employment is growing above trend) would be followed by a similar period when growth is below trend. Thus, on average, the economy grows at a relatively stable trend rate, but it cycles above and below the trend.

On the other hand, followers of RBC claim that the actual growth rate over time follows a 'random walk with drift'. In other words, real GDP tends to grow at a relatively stable rate, say, three per cent, which is the 'drift', but over each year it can deviate from that trend in a 'random walk' (random deviation from the trend). A period above trend is no more likely to be followed by a period below trend as it is to be followed by a period above trend. Since they reject the notion of a conventional 'cycle' altogether, it is somewhat ironic that their approach is called 'Real Business Cycle'.

29.4 New Keynesian Economics

Introduction

Over the past three decades another mainstream school has developed, in part in a reaction against some of the more extreme assumptions underpinning the New Classical and Real Business Cycle approaches. New Keynesian Economics (NKE) attempts to resurrect some of the ideas adopted in the post-war Neoclassical Synthesis approach.

Most importantly, NKE rejects the assumption of continuous market clearing. While adherents do accept the metaphor of an invisible hand guiding the economy toward equilibrium, they believe there are a number of impediments or rigidities that slow down the market adjustment processes.

For example, if wages and prices are not perfectly flexible then the market cannot quickly produce prices that will equilibrate the quantity supplied with the quantity demanded. An example of wage rigidity would be a legislated minimum wage law, since the wage cannot fall below the legal minimum, and unemployed workers cannot offer to work at a lower, market clearing wage.

The NKE invokes Keynes' name, attributing to him the argument that wage rigidity is the cause of unemployment. However, as we discussed in Chapter 12, Keynes actually argued that unemployment results from an insufficiency of aggregate demand. He went on to suggest in Chapter 19 of the *General Theory* that falling wages could make matters worse by lowering effective demand, since workers with lower money wages, all else being equal, would have less real income to spend and the expectations of entrepreneurs would be depressed. Keynes blamed the use of money, which allows an agent to save in liquid form, rather than wage rigidity for unemployment.

Thus, while both NKE and Keynes reach a similar conclusion that markets do not generate continuous full employment, their explanations are quite different.

Examples of price and wage inflexibility

NKE has tried to come up with 'rational' explanations for rigid wages and prices. For example, NKE theorists have assumed that it is costly to change prices frequently, which leads to quantity rationing rather than price rationing.

Take the example of a restaurant that incurs the costs of printing menus. It can be more profitable to hold prices steady, even in the face of rising restaurant costs, in order to avoid the 'menu costs' of raising them. What the restaurant would gain by higher prices for meals is more than offset by the cost of printing new menus. Also, holding prices steady in these circumstances may promote consumer 'goodwill' and better sales.

While NKE often adopts the assumption of rational expectations, its practitioners sometimes assume that there are 'information asymmetries' that interfere with price adjustment to clear markets. While there are many types of asymmetry, a couple of examples will be sufficient.

NKE acknowledges that it might promote good relations with workers if wages were to be held steady in the face of deflationary pressures caused by unemployment. Although in times of high unemployment the employer could find workers willing to replace existing staff at a lower wage, it might be more profitable for the firm to resist the temptation to cut wages. The savings from lower wages paid could be more than offset by higher costs due to recruiting and training new, lower-paid workers. There also could be information asymmetries; that is, some unemployed workers may be willing to work for lower wages but might be less productive, and the employer might find it hard to identify them, and so on. Also, the morale of current workers might suffer if wages are cut, which reduces effort/ productivity so that shirking on the job might increase.

Thus, NKE recognises that higher wages make workers happier and healthier, which can make them more productive, and hence it is more profitable to keep wages high. Worker productivity is thus linked to wages paid.

A second information asymmetry relates to the risk of a loan default being known to the borrower but not to the lender. There are 'good' borrowers who repay loans and 'bad' borrowers who will not, but the bank cannot tell them apart.

As the bank raises interest rates to compensate for default risk, some 'good' borrowers will decide not to borrow because, given higher rates, the servicing costs of the loan have risen. However, since the 'bad' borrowers are going to default, they remain in the 'pool' no matter how high the rate rises. They then default when the payments come due.

It would not be sensible for the bank to try to clear the market for loans by raising rates until the quantity of loans supplied equals the demand, because loan defaults will not fall since it is the good borrowers that leave the pool. The problem here is one of 'adverse selection', because using just price rationing will lead to the bank selecting from a worse pool of borrowers.

It is 'optimal' to leave interest rates lower, but restrict the quantity of loans. In other words, the bank uses 'quantity rationing' rather than 'price rationing'. However, there is an unsatisfied residual of 'credit rationed' borrowers.

There can also be 'adverse incentive' problems. As rates rise, some borrowers may choose riskier projects. This is because they believe there is a 'risk return' trade-off: with riskier projects having a higher expected return to compensate for the greater risk. Again, the assumption is that the borrowers know which projects are riskier, but the bank does not and so it cannot devise a contract that would prevent borrowers from switching projects after a loan is made.

Again, if the bank raises rates to allocate loans by price, it will lend to a riskier pool of borrowers (this time because the borrowers who remain in the pool switch to riskier projects). Thus, again, it is better for the banks to use quantity rationing while keeping rates low, due to the adverse incentive problem. Hence low rates reduce both 'adverse selection' problems (making loans to riskier borrowers) as well as 'adverse incentive' problems (borrowers switching to riskier projects). Thus, interest rate inflexibility means that the market for loans does not clear.

The role of policy

In the context of rigid wages and prices, NKE sees a role for active policy. For example, long-term labour contracts can create wage rigidities and prevent continuous market clearing, which leads to quantity rationing. These contracts can be rational for reasons noted above: to provide security, reduce hiring and retention costs, and raise worker morale.

Policy can reduce the unemployment that is alleged to result from wage rigidity. For example, assume that workers and employers negotiate a nominal wage over a two-year labour contract period. We can presume that workers hold 'rational expectations' about inflation, which influence their negotiation strategy.

However, they do not have perfect foresight, so their expectations can be wrong due to random shocks to prices. If those happen to be negative shocks, inflation will be less than foreseen when the contracts were signed. That means that the real wage will be higher than firms expected due to the predetermined (sticky) nominal money wage.

Firms cannot lower nominal wages, so they must instead lay off workers, whose marginal productivity is less than the real wage in this neoclassical model of the labour market. Unemployment results. Policy will be effective if it can raise nominal demand, pushing up inflation to its expected rate. Government could use either fiscal or monetary policy.

Note that policy has real effects even if expectations were formed rationally. Policy can be changed more frequently (that is, during the contract period) than labour contracts.

More generally, discretionary policy can attenuate the business cycle if the assumption of continuous market clearing is dropped. While the assumption of rational expectations appears to lead to the conclusion of policy ineffectiveness, NKE demonstrates that the critical assumption is continuous market clearing. If markets do not continuously clear, there is room for effective policy.

Gregory Mankiw (2016), author of a mainstream textbook, argues that while RBC (and NC) might be appealing due to its "internal consistency", it is not "externally consistent". In other words, RBC's assumptions (that is, rational expectations, neutral money, and continuous market clearing) are internally consistent with a rigorous neoclassical approach, but they do not provide an empirically plausible explanation of the real world economy. By contrast, he argues, NKE is not internally consistent but fits the facts better. It is hard to believe that either (unanticipated) 'money shocks' in the case of NC, or 'technology shocks' in the case of RBC can explain the real world business cycle.

On the other hand, the explanations given for wage and price rigidity are somewhat *ad hoc*. They are not rigorously grounded in neoclassical micro foundations. But some wage and price rigidity is observed and even anticipated monetary policy changes can have real effects. If fluctuations of employment are not truly 'optimal' and if policy is not completely impotent, then it is prudent to use discretionary policy to address, say high unemployment. It would be dangerous to allow unemployment to rise to double digits, despite the claim from rigorous RBC and the NC models that higher unemployment must be 'optimal', since they critically rely on the continuous market clearing assumption.

29.5 Modern Heterodox Schools of Thought

Introduction

Throughout this text, we have presented different theoretical aspects of the main heterodox schools of thought and frequently refer to the history of the development of these schools.

In this section we will focus on contrasting heterodoxy with the mainstream macroeconomic theory and policy outlined previously in this chapter. While modern heterodox schools of thought do have some differences with one another, they share general features that are quite different from the approaches examined above.

Let us first provide a short list of the main distinguishing features generally shared by orthodoxy that contrast sharply with positions taken by heterodox macroeconomics:

- Orthodoxy is based on methodological individualism in which decision making by individuals is analysed taking prices as signals. Equilibrium analysis then shows how market forces drive the economy towards equilibrium where all markets clear.
- Marginal productivity theory determines the reward of each factor of production and hence the distribution of income.
- Say's Law holds, with full employment equilibrium, at least in the long run.

- The loanable funds market determines interest rates at the point where the supply of saving equals the demand for funds resulting from borrowing for investment.
- Perfect competition is assumed, so that there are many buyers and sellers in the market, none of whom can influence the price.
- Money is neutral, at least in the long run, and financial markets are efficient so that finance doesn't really matter.
- Logical time and rational expectations ensure the consistency of expectations and outcomes.

We now turn to alternative heterodox perspectives on these topics.

Method: the notion of equilibrium and locus of analysis

In neoclassical theory, equilibrium is identified with market clearing. In heterodox thought, there are two main alternatives. Some writers reject the neoclassical notion of equilibrium altogether and others adopt a modified approach to equilibrium.

1. **Disequilibrium or Nonequilibrium approaches:** This heterodox approach stresses imperfect foresight, uncertainty, and the whirlwinds of speculation outlined in Chapter 12 of the *General Theory*. Uncertainty implies that an individual cannot optimise or maximise. At best, people operate with bounded rationality and adopt rules of thumb. Expectation formation is endogenous and subject to cumulative causation.

   The economy evolves, so that economics is more akin to the study of biological evolution. We have some understanding of where the economy has been but we cannot know where it is going. Analysts must take an historical approach to studying the economy since the future is non-deterministic.

   Economics must take account of the dynamic nature of an economy that is subject to cyclical behaviour. The evolution through time need not be tranquil, since financial fragility can grow until a crisis is precipitated. Most importantly, market forces may not be stabilising. Positive feedback might even move the economy further and further from a neoclassical market clearing equilibrium.

2. **Heterodox equilibrium** approaches: Recall that Keynes defined equilibrium as the point where the D curve crosses the Z curve (see Chapter 13), the point of effective demand. A change of long run expectations can induce changes of spending, which then affect employment and income through the multiplier.

   Essentially, we can analyse the economy through time as a series of snapshots taken at different points of time, each of which shows a different equilibrium point where D = Z at some level of employment. There is no reason to believe that any of these positions of rest coincides with full employment. Thus, this concept of equilibrium is quite different from the neoclassical notion of market clearing.

   Some early Cambridge theories of economic growth and distribution extended Keynes' work on effective demand from D/Z equilibrium that occurs at a point of time to the growth of the economy over time (see, for example, Domar, 1946).

   It is even less likely that effective demand would be consistent with full employment over time. Not only are there multiplier effects on aggregate demand induced by investment, but also investment in plant and equipment increases capacity.

   The question is whether the demand side effects of investment would be equivalent to the supply side effects. If demand side effects are larger, the economy would suffer from insufficient capacity and hence tend toward explosive growth. On the other hand, if supply side effects are larger, the economy would always have too much capacity resulting in idle capacity that lowers demand and capacity utilisation, via lower investment. The equilibrium conditions that enable the economy to grow steadily through time, albeit not necessarily at full employment, are restrictive.

   The point to emphasise here is that some heterodox economists follow Keynes in using an equilibrium methodology. Again, this is not in the sense of market clearing, but rather in the sense of decision formation in conditions of uncertainty. Entrepreneurs must decide how many workers to hire in order to produce the amount of output they think they can sell at profit. Given expectations, they do not want to hire more.

The resulting quantity of employment and output generated is an equilibrium in the sense that it is consistent with expectations, but it is not necessarily associated with a level of output and employment where Z and D are equal. This is often described as a 'state of rest'. No one would choose to do anything differently, given expectations. To be sure, if expectations were higher or lower, the economy would come to a different 'state of rest' equilibrium.

Heterodoxy also rejects the neoclassical methodology based on rational, maximising, self-interested individuals whose behaviour can then be aggregated up to represent the operation of the economy as a whole. They reject acquiescence to a supposed 'invisible hand' while affirming the importance of collective action.

The economy is seen as a system of social control, not one of independent individuals making decisions based on signals provided by autonomous markets. Institutions structure markets and are largely responsible for allocations. Heterodoxy favours pluralism and democracy. Outcomes are not the result of economic laws, but are subject to human decisions and can be changed.

Critical realists emphasise that there are deep structures in the economy that are hard to identify. The analyst moves from stylised facts to theory to analyse underlying relations and structures. The scope for prediction is limited. We must begin with observation and build on realistic abstractions.

The **Pragmatic Institutional** approach emphasises a multidisciplinary approach to problem recognition and problem solving as the goal of the economist. Institutionalists reject the notion of a natural economy. The economy consists of human institutions that can be changed.

To sum up, heterodox economists have used a different concept of equilibrium: one that need not imply market clearing but one that is consistent with the operation of a capitalist system.

Alternative approaches to distribution

In neoclassical theory, the distribution of incomes follows the marginal productivity of the factor inputs and is thus determined within the production process. In most heterodox approaches, distribution cannot be explained within the production sphere; it is separately determined by power, institutions, bargaining, monetary and fiscal policy, and so on.

For example, the data on gender wage differentials demonstrate beyond any doubt that males earn more than females across occupations, after controlling for all conceivable factors that might affect wages. Similar differentials are found for country-specific racial, ethnic, and religious groups.

Further, given the complexity of modern production processes, in which thousands of workers perform myriad tasks across space and time to produce differentiated products sold in various markets, it would be impossible for a firm to calculate the marginal productivity of any individual worker.

All production is social in the sense that it is usually produced by teams and always utilises the know-how created by past generations of humans. There is, quite simply, no such phenomenon as an individual's contribution to production that can be readily measured and rewarded.

For these reasons, heterodoxy has always taken a different approach to the study of distribution. Rather than looking to individual differences (such as productivity) to explain distribution, heterodoxy has often looked at aggregates to explain distribution; bargaining by groups (such as labour unions), discrimination (race, gender, ethnicity), functional categories (firms versus households), and social classes (workers versus capitalists).

In the remainder of this subsection we look at heterodox work in the last two areas (firms and capitalists versus workers and households). This work largely begins with the recognition that there is a fundamental power differential, with capitalists and firms making the decisions that largely determine the distribution available to workers and hence household consumers.

Early followers of Keynes, such as Joan Robinson, worked out how changes in variables, including the distribution between wages and profits, and different propensities to consume of different groups of consumers affected growth.

This is related to Kalecki's (1971) dictum: workers spend what they get; capitalists get what they spend. As we saw in Chapter 25, Kalecki showed that the more capitalists spend (on investment or their own consumption) the higher their profits. Thus, their total spending creates their collective income. By contrast, workers rely on hiring decisions made by firms. Worker spending on consumption is largely determined by their wage income.

Bargaining by workers can affect the distribution of real wages among workers, but has no impact on the distribution between workers and capitalists (that is, the division between wages and profits).

How do we know that, even in the short run, capitalists will secure profits? In neoclassical theory, each factor's income depends on its individual marginal product. Heterodoxy rejects this. Marx had already tackled this question in order to determine the conditions of simple reproduction (keeping output constant) and expanded reproduction (to allow for growth), each of which requires a division of output between investment goods and wage (or consumption) goods to ensure that profits are realised.

In the simple version of reproduction, capitalists merely produce enough to replace the worn-out means of production plus enough to give workers the necessary wage to achieve the socially accepted living standard. Expected profits must be high enough to induce enough production to provide for consumption and replacement investment that will enable the economy to continue to produce its current level of output. In the expanded model of reproduction, production is high enough to add more means of production so as to enable growth to occur.

There are two further considerations that come out of this research. First, the distribution between wages and profits is not based on productivities *per se*, but it is necessary to generate a surplus (surplus value) over and above workers' subsistence, otherwise profit cannot be realised. However, there will be no profits in this period if there is no investment or capitalist consumption in this period.

Thus, the two critical variables are the capacity to create a surplus and capitalist expenditure in the form of investment (or the share of investment in output), and the propensity to consume out of profit.

Pasinetti (1962) showed that if workers do save, they get some profit income. The wage share is unaffected, but worker saving does affect the overall income share going to workers.

In neoclassical theory, the marginal product of capital determines the profit rate received by capitalist owners of capital. Keynes rejected this, arguing that capital gets a return to scarcity, with the interest rate setting the standard.

Further, Sraffa (1960) showed that marginal productivity theory is logically flawed. Neoclassicals want the marginal product of capital to determine the rate of profit; but we cannot value heterogeneous capital unless we know the rate of profit (since the rate of profit depends on both the flow of profits and the value of the capital stock).

Marginal productivity theory can establish the rate of profit that would be implied in the existing price of the capital stock, but it cannot explain the rate of profit. Sraffa (1960) showed that if capital is a produced good, then we cannot value capital (or know aggregate income) until distribution and the rate of profit are given. A change of distribution will change the value of capital.

Further, since the marginal product of labour must cover not only the wage, but also the profit the capitalist receives on wages advanced, marginal productivity theory also breaks down for labour. So, we cannot use marginal productivity theory to explain either profits or wages; and we cannot sum up marginal products to get aggregate income since that depends on distribution.[1]

The second development that followed from Marx's reproduction schemes was Sraffa's search to see if it is possible to find equilibrium prices which are not designed to clear markets but to reproduce the system, covering costs and giving equal profit rates to all forms of production.

Essentially it is a cost plus mark-up, input-output model. Sraffa takes the rate of profit as given, along with wages (to provide the socially necessary standard of living), and then derives the system of prices that ensure equal rate of profit. In general, a change of distribution between wages and profit will lead to a different system of prices (that will reproduce the system).

Individual firm-level mark-ups are determined by pricing power, which distributes profits among firms from the aggregate pool of profits generated by capitalist consumption (see the discussion of perfect competition below).

Similarly, competition among workers distributes the aggregate of wage goods. The analogue to firm pricing power is the bargaining strength of worker groups, which does not determine the aggregate wage but rather the distribution of the aggregate among workers.

Heterodox macroeconomists begin with the determination of output and employment as a whole, that is, Keynes' point of effective demand, and outline the decision processes that generate that outcome. In the capitalist economy, the locus of power in the private sector rests with capitalists and firms whose decisions determine not only the level of output and employment as a whole, but also the division of output between capitalists and workers. The degree of monopoly power then distributes the aggregate pool of profits among capitalists. Competition among worker groups distributes the residual aggregate wages among workers.

Say's Law

Neoclassical economists adopt the strong version of Say's Law in that not only does supply create demand, but flexible wages ensure that supply equals demand at full employment (see also Chapter 12). Heterodox economists reject both conclusions. As Keynes explained, expected effective demand determines employment and output and there is no reason to suppose that this is consistent with full employment. Moreover, the supply of output does not create equivalent demand if there is a way in which purchasing power can be stored without demanding current output. In other words, if income can be saved in the form of money, then the creation of income today does not mean that it will be spent today.

As we discussed in Chapter 12, Keynes insisted that the decision to save is a two-step affair: first one decides not to consume and second in what form to save. Only if saving were limited to ordering goods and services for future consumption would the decision to save automatically lead to a demand for production, ensuring that Say's Law would hold. Otherwise, 'supply' will not create a 'demand' for output.

Another way of stating this point is via Keynes argument that so long as we have two types of aggregate demand, with one type that is not a function of current income, then there is no reason to expect that increasing income will raise spending by the same amount.

Loanable funds versus liquidity preference

In orthodoxy, the interest rate is typically presumed to be determined in the market for loanable funds, with saving representing the supply and investment creating the demand. Following Keynes, heterodox economists argue that because investment creates saving (or more generally, injections create leakages) the saving cannot pre-exist investment. It is thus logically impossible for saving to be the source of finance for investment. Further, heterodoxy recognises Keynes' 'paradox of thrift'. The attempt by households to increase total saving by reducing consumption does not increase saving, other things being equal, but rather reduces income. This is related to the rejection of Say's Law which was discussed in the previous section.

Many heterodox economists follow Keynes in substituting a liquidity preference theory of interest rate determination for the loanable funds theory. This is developed in Chapter 17 of Keynes' *General Theory*.

The receipt of interest payments is a reward for parting with liquidity. It is related not to the first step in the saving decision (that is, the decision not to spend), but rather to the second step, namely the form in which saving is to be held. To induce one to save in a less liquid form, the saver demands a return – interest – with the rate of return depending inversely on the asset's liquidity (see also Chapter 12).

Note that there is a continuum of liquidity, ranging from the most liquid assets (cash, reserves, demand deposits) through quite liquid assets (long-term government and corporate bonds and mortgage-backed securities) and to relatively illiquid assets (plant and capital equipment).

The return on cash is zero, but the return on bank reserves is set by policy (the overnight rate target of the central bank). For this reason, some heterodox economists argue that the interest rate is determined 'exogenously' by policy. This is strictly true for the overnight rate.

Other interest rates have complex determinants but are influenced by expectations of future monetary policy. If the central bank is expected to raise interest rates in the future, the rate on longer term assets (such as 10 year or 30 year treasuries) will rise today because the future rate hikes will cause prices of such assets to fall at that time (rates and prices on bonds move inversely). This is called the **expectations theory of interest rates**.

Thus, the loanable funds theory is rejected. Interest rates are not determined by saving and investment decisions but rather are determined in asset markets by liquidity preference as well as monetary policy decisions.

Imperfect competition

Many heterodox traditions have long rejected perfect competition as the base case. Instead it is recognised that many of the most important prices are administered by firms with market power. Rather than accepting prices as given by market supply and demand, firms are seen as setting prices as a mark-up over per-unit costs.

While there are several different approaches to mark-up pricing, the basic idea is that firms need to cover production costs (largely, wages and intermediate goods) plus a mark-up to cover rent, interest, taxes, certain types of discretionary overheads, and profits. This was outlined in more detail in Chapter 16. The degree of competition (or Kalecki's degree of monopoly) contributes to the determination of the size of the mark-up.

The imposition of mark-up pricing is inconsistent with price flexibility which is alleged to lead to market clearing. If firms allowed prices to fall, they would endanger their financial viability. Further, in a situation of oligopoly market structure, price cuts could trigger a price war with ongoing price cuts being designed to maintain/restore market share.

Flexible prices are especially dangerous for indebted firms. If they cannot cover loan commitments, they can be forced into bankruptcy. Indeed, firms without price setting power have limited access to credit markets. In short, the capacity to ensure that prices are sufficient to generate revenues to cover loan payments is critical to obtaining credit.

Some heterodox economists distinguish between the mark-up set at the microeconomic level by firms and an aggregate mark-up at the macro level.

While prices are administered by firms, there are competitive and other constraints on setting high prices which include: (a) consumers who substitute away from the particular product; (b) competitors who have an incentive to enter the industry; (c) government intervention if price gouging is thought to be occurring; (d) labour unions (workers) who may object to the pricing behaviour of the firms for which they work; and (e) depressed macroeconomic conditions that prevent the realisation of the desired mark-ups, so that firms may have to slash prices to sell output.

There is another reason that the mark-ups desired at the micro level depend on macro-level spending. The Kalecki equation links aggregate capitalist spending to aggregate profits. Thus, holding all else equal, for profits to increase, capitalist spending must rise. Aggregate profits are determined by total capitalist spending (capitalists get what they spend).

This is without reference to micro-level pricing behaviour. If capitalist spending is zero (in the simple model) then aggregate profits are zero. No amount of pricing power can affect aggregate profits in such a condition.

How can we reconcile the microeconomic outcomes with the macroeconomic conditions? Given that macro profits are determined by investment in the simple model, micro-level pricing power determines the **distribution** of profits among capitalists. Those that have more pricing power can get their desired mark-up; those with less market power set lower mark-ups and might even make losses.

Obviously, outcomes affect expectations of the future, which in turn, affect investment and thus profits in the future. This means that theorists must be careful in aggregating up from individual decisions, whether they are those of consumers or capitalists. The example of a zero capitalist spending environment demonstrates that administered pricing *per se* does not create profits.

Treatment of money, time and expectations

The key components of the 'fundamentalist' Keynesian approach to money include:

1. The presence of true uncertainty.
2. The existence of market institutions, notably money, money contracts, and sticky money wages.

BOX 29.1 WHY DIDN'T KEYNES REJECT PERFECT COMPETITION IN THE GENERAL THEORY?

Keynes did not reject the neoclassical assumption of perfect competition. His strategy was usually to accept as much common ground as possible with (neo)Classical thinkers and then to reject only those assumptions that he believed to be critical to the point he was trying to make. His followers are divided on the wisdom of such a strategy.

Some, especially those who have read the work of Kalecki, insist that this was a mistake. It would have been much simpler and more realistic to simply reject the notion of perfect competition and instead include different degrees of competition across industries.

Even orthodox economists accept the conclusion that in the absence of perfect competition, markets do not clear because the optimal strategy for firms with market power is to restrict output in order to control price. Prices and wages, then, are not set to clear markets.

However, Keynesian fundamentalists argue that the degree of competition is irrelevant to his model. All that is needed is that output and employment are both functions of expected effective demand. Then Say's Law is broken.

Furthermore, as Keynes argued, wages and prices *are sticky*: people will not use money contracts nor hold money as a temporary store of purchasing power, unless prices are relatively sticky. They have to have some confidence that flow supply prices do not vary much. Flow supply prices are the prices that are required to induce producers to produce additional output. These prices must be sticky, but at a price high enough to cover costs with a margin.

On the other hand, spot prices might be more flexible. These are the current prices at which output can be sold out of inventory. They may not be high enough to induce more production. Only if spot prices rise higher than flow supply prices will more output be forthcoming. So, while spot prices can be relatively flexible, flow supply prices need to be fairly rigid to maintain conditions that are conducive to production.

Also Keynes argued, being able to hold money and to write contracts, along with stickiness of price, allows individuals to delay decisions to the future. If money prices could fluctuate wildly, it would be too risky to hold money in order to delay decisions since one would not know how much the money held would be worth in terms of resources or output. A 'perfect' market in labour with flexible wages would not be compatible with a monetary economy. Money could not serve as the link between the present and the future, unless there is some stickiness.

3. Money serves both as a medium of exchange and a store of value. In addition, money must have two essential properties: (a) a negligible elasticity of production, so that minimal labour is required in its production. This means that when the demand for money to hoard rises, and aggregate demand falls, total employment falls; (b) a negligible elasticity of substitution so that when the demand for money to hoard rises, this cannot be satisfied by other goods or assets.

The following propositions result from the key components of the fundamentalist Keynesian approach:

1. If uncertainty exists, then even in the long run, there may be no equilibrium at full employment.
2. The stickiness of the money wage is essential if money is to play its peculiar role of acting as the unit of account in which contracts are written and also serving as a store of value.
3. If wages and prices are fully flexible, then money will not matter in the long run; but something else would take its place.

Being able to hold money and to write contracts in money terms, along with stickiness of prices, allows individuals to delay decisions to the future. A perfect labour market with flexible wages would not be compatible with a monetary economy. We hold money because we do not trust our guesses about the future and it provides us with a reliable store of value so long as wages and prices are relatively sticky in terms of money. (Note that this does not require that inflation is zero but only that we can with reasonable confidence project money's value into the future.)

The existence of true uncertainty makes it impossible to optimise or maximise because we simply do not know the future. Inevitably we undertake decisions by exercising bounded rationality. This might lead us to conclude that the safest thing to do is to fall back on opinions and expectations of others. We use conventions and weight of argument to decide what to do, without knowing if this is optimal.

Some post-Keynesians, in particular Paul Davidson, argue that the future is non-ergodic. What this means is that even if we had all the probability distributions for past outcomes, these cannot be applied to the future because the probabilities for the future could be, and probably will be, different from those calculated from the past.

The easiest way to think about this is to remember that the future is yet to be created, and no matter what humans have done in the past, they could choose to do something different in the future. To some degree we could say that we know quite a lot about the past and even about the present, but we simply cannot know the future, at least, until it occurs.

Minsky had a slightly different view on uncertainty. The uncertainty we face is over the correctness of our model of the way the world works.[2] The uncertainty is not just about the data and probabilities, but rather about the correct model of the world.

We know our model can be wrong. Hence, we build in margins of error and the best margin of safety is to hold highly liquid assets like cash. It might turn out that our model of the world is wrong, but holding liquid assets provides a cushion of safety.

Joan Robinson used to joke that time exists to keep everything from happening at once. This does capture a key difference between heterodoxy, which uses the concept of historical time where everything cannot happen at once, and orthodoxy, which uses the notion of logical time in its rigorous models. In logical time, one can go backward and forward in time to re-contract in order to get the best deals.

In the Walrasian auctioneer models, the auctioneer takes bids and offers but nothing happens until the equilibrium price list for all tradeable goods and services is settled. Once equilibrium prices are established, then all exchanges take place.

This ensures there is no false price trading at disequilibrium prices. In such models, there is no room for money, since there is no reason to postpone purchases to see whether a better opportunity might come along. And there is no uncertainty because the equilibrium prices reveal all the information that is needed. Effectively, time collapses to the instant during which the exchanges take place. All exchange reduces to barter.

By contrast, in heterodoxy, money matters:

- We need money to spend, as Clower explains. Money buys goods and goods buy money but goods do not buy goods.
- We measure success in terms of money. Marx argues that capitalist production takes the form of *M-C-P-C'-M'*: money (*M*) is advanced now to buy inputs (C) and produce (P) commodities (C') with an expectation that these can be sold for more money (*M'*) later.
- We can hold money against an uncertain future.

Money is needed to start the production process, to finance spending by both firms and consumers. However, money also allows one to defer decisions; it is a safe repository that breaks the link between income and spending, nullifying Say's Law. When uncertainty rises, the demand for liquidity simultaneously rises and there is a growing gap between the supply and demand for goods and services. Money is never neutral, although in an important sense it matters more in troubled times.

In addition to these functions of money, MMT adds the link between money and sovereign power, which is largely lacking in other heterodox approaches to money, as well as in orthodoxy. The state also needs to finance its spending and plays an important role in organising the monetary system to move resources to the public sphere. Money cannot be neutral for the state either, since currency sovereignty is critically important to ensure that the state is not financially constrained.

The neoclassical notion of Ricardian Equivalence cannot hold in the case of a sovereign currency. If a government spends more or taxes less, it would be irrational for the private sector to cut its own spending in anticipation of future tax hikes.

Sovereign governments do not need to raise taxes in the future to pay for spending (or tax cuts) today. And indeed, they do not do so in the real world. We observe that the normal situation for most sovereign governments is to run nearly continuous deficits and rarely, if ever, pay down a significant portion of their debt. This is for reasons discussed in Chapter 21.

What matters, in any case, is to run the economy near to continuous full employment of its resources. Any labour resources not used this year cannot be stockpiled for future use, nor can those labour resources used this year be somehow paid for by taxes later. Any resources mobilised this year for use in the public sector are paid for immediately, by cutting cheques or by marking up bank accounts.

It would be irrational for the private sector to react to greater receipts from government spending by cutting back household and business spending. Rather, it is far more rational to increase private spending alongside rising public spending.

These ideas underpin the concept of the Keynesian multiplier. All plausible empirical estimates demonstrate that there is a multiplier effect, not a Ricardian effect, of government spending.

MMT also teaches that central banks always operate with an overnight interest rate target, which is consistent with heterodox as well as NMC thinking.

This implies that central banks must coordinate operations with the fiscal authorities, offsetting the effects of fiscal operations on reserves (either by accommodating bank demand or by paying interest on reserves).[3] However, all else being equal, budget deficits increase excess reserves that would place downward pressure on overnight rates, which is precisely the opposite result to that implied by the neoclassical crowding-out theory.

Crowding-out theory assumes that there is competition between government borrowing and private borrowing for a limited supply of loanable funds. This completely misunderstands government finance, as well as monetary policy operations. If interest rates do go up in the presence of budget deficits, it is because the central bank has raised its target rate.

Conclusion

In this chapter we continued our discussion of the history of economic thought by delving more deeply into the main modern schools of thought. We began with the orthodox approaches, including New Classical, Real Business Cycle, and New Keynesian. We explained how each developed as a response to perceived weaknesses in the older Monetarist and IS-LM approaches, and then looked at the weaknesses in each of these approaches. In the next chapter we will study the current 'synthesis' that attempts to integrate parts of each of these three schools.

We next discussed in detail the shared fundamental building blocks of the modern heterodox schools of thought, and briefly contrasted these with the foundations of mainstream, orthodoxy.

References

Domar, E.D. (1946) "Capital Expansion, Rate of Growth, and Employment", *Econometrica*, 14(2), April, 137–47.

Friedman, M. (1968) "The Role of Monetary Policy", *The American Economic Review*, 58(1), March, 1–17.

Friedman, M. (1970) "The Counter-Revolution in Monetary Theory", IEA Occasional Paper, No. 33, Institute of Economic Affairs, London.

Kalecki, M. (1971) *Selected Essays on the Dynamics of the Capitalist Economy*, Cambridge: Cambridge University Press, pp.78–9.

Lucas, R.E. (1972) "Expectations and the Neutrality of Money", *Journal of Economic Theory*, 4(2), 103–24.

Mankiw, N.G. (2016) *Macroeconomics*, 9th edn, New York: Worth.

Pasinetti, L. (1962) "Rate of profit and Income Distribution in Relation to the Rate of Economic Growth", *Review of Economic Studies*, XXIX (4), October, 267–279.

Sraffa, P. (1960) *Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory*, Cambridge: Cambridge University Press.

Endnotes

1. Aggregate income does not depend on distribution unless capital is homogenous with what Marx called a constant organic composition of capital; this requires essentially the same ratio of hours of labour employed directly in production to the hours of labour embodied in the capital used. On the other hand, Keynes has no such problem in the exposition of the *General Theory* since he measures everything in labour hours and a nominal wage unit.
2. Minsky credited Lucas and the NC approach for recognising that agents in the model have a model of the model, but he rejected the NC presumption that the model held by agents is the correct one. He argued that in the real world, everybody knows the model held is wrong. This is why it is prudent to hedge one's bets by building in margins of safety.
3. If interest is not paid on reserves, the central bank ensures excess reserves are kept near zero while offering treasury bonds through open market operations as the interest earning alternative. If the central bank pays a rate of interest on reserves equal to the target rate, there is no need to drain excess reserves by offering bonds.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

30 THE NEW MONETARY CONSENSUS IN MACROECONOMICS

Chapter Outline

Learning Objective

- Appreciate how elements from the different orthodox schools of thought have been fused into the dominant mainstream New Monetary Consensus in macroeconomics.

30.1 Introduction

Since the 1990s an attempt has been made to synthesise the various strands of mainstream macroeconomics. It is referred to as either the New Monetary Consensus or the New Macro Consensus (NMC). This model is an integration of ideas taken from the New Classical (NC), Real Business Cycle (RBC), and New Keynesian (NK) approaches. It usually adopts rational expectations, but also presumes that wages and prices can be sticky, as in the New Keynesian model. The main focus of this approach is monetary policy, with most followers relegating fiscal policy to the background, although some are willing to advocate fiscal stimulus in a deep recession. Economic shocks can come from money 'surprises' (NC) but also can come from the real sector (RBC).

The old Monetarist notion that central banks do, and should, control the money supply has been abandoned. Instead, it is recognised that monetary policy targets an overnight interest rate. Many central banks use that target to try to achieve an explicit inflation target, although others adopt a less rigid approach, preferring to allow inflation to move within some band. Still, the belief is that monetary policy is a powerful tool to pursue price stability, and economic stability more generally.

Some NMC advocates support a nominal GDP growth target. Nominal GDP growth equals the sum of real GDP growth plus the inflation rate. This is supposed to be achieved by setting the nominal interest rate in response to a 'demand gap', that is, the deviation of actual output from the output that would be generated at full capacity use, and an 'inflation gap' (deviation of inflation from the desired rate). A 'Taylor Rule' is often recommended as a guide to interest rate setting (see the Appendix to this chapter). According to the Taylor Rule, the central bank should raise interest rates if inflation is higher than desired, and lower the interest rate target if inflation is too low.

NMC economists also adopt a Dynamic Stochastic General Equilibrium model (DSGE) for studying the economy and in formulating central bank policy. The approach is dynamic in that it models growth and the cyclical behaviour of the economy. The model is stochastic since it allows for random errors or shocks that temporarily disturb the economy, and it takes a general equilibrium approach based on neoclassical micro foundations of decision making by rational, utility-maximising, forward-looking individuals. It is not subject to the Lucas critique (see Chapter 29) since individuals are assumed to hold rational expectations. In fact, using the term 'individuals' is a misnomer, because to avoid aggregation issues across all individuals, these models typically use a 'representative agent' (a single consumer or household or firm) and assume all individuals behave like this agent.

Prior to the Global Financial Crisis (GFC), the NMC had become the dominant approach to orthodox macroeconomics theorising during and after the 1990s – a period that became known as the 'Great Moderation'. The NMC retained its dominant status after the GFC even though NMC theorists had done no better than other orthodox economists in foreseeing the crisis, and even though the general forecasting accuracy of their models was very poor. In defence of their poor track record, the NMC theorists claimed that their approach was more easily adapted to explain the crisis *ex post* – after it had happened. This is because it was obvious that neither a 'monetary shock' (as in Monetarism Mark I or Mark II) nor a 'real shock' (as in RBC) could be blamed for the crisis. It was recognised that the financial sector had something to do with the origins of the crisis, and that insufficient aggregate demand prolonged the downturn. As such, the goal was to return to a more 'Keynesian' approach, and to include the financial sector in the analysis.

It was also recognised that the old 'Keynesian' workhorse model, IS-LM (see Chapter 28) would not do. The IS-LM approach was based on an assumption that the central bank could control the money supply. As such, the LM curve had to be discarded and replaced with interest rate setting by the central bank. Further, the IS curve had to be modified to explicitly recognise that aggregate demand need not be consistent with full employment. Still, NMC retained many of the assumptions from the New Classical, Real Business Cycle and New Keynesian toolkits.

In this chapter, we will examine the NMC and assess its ability to incorporate important real world phenomena such as involuntary unemployment and the financial sector. The treatment is necessarily somewhat advanced.

30.2 Components of the NMC theory

NMC adopts many of the traditional neoclassical assumptions, as modified by New Keynesian theory:

1. Both micro and macro outcomes are derived from the intertemporal decision making of rational individuals. That is, individuals are assumed to maximise utility at each point in time for all time. Thus, the basic behavioural presumptions of the NC and RBC schools are used as the starting point.
2. The economy moves toward equilibrium in the long run; however, rigidities (mostly 'sticky wages') prevent markets from clearing in the short run – the NK approach.
3. These short-run deviations from equilibrium are best addressed by monetary policy. Monetary policy sets the interest rate to hit an inflation target.
4. The NMC adopts the unemployed buffer stock approach, by which 'optimal' employment is consistent with the level of output that would occur if wages and prices were perfectly flexible – which equals the non-accelerating inflation rate of unemployment (NAIRU). Full employment is not defined in terms of a number of jobs that will satisfy the preferences of workers for hours of employment, but rather the unemployment rate that is consistent with stable inflation. Rigid wages and prices can move unemployment above the NAIRU level in response to shocks (from either the demand or supply side – see below).
5. Unlike traditional 'Keynesian' structural models used in the early post-Second World War period to guide policy design and implementation, the NMC policy models are grounded in neoclassical microfoundations (see point 1 above), with rational expectations presumed (see below).
6. NMC takes the 'Lucas Critique' into account while constructing models by making expectations endogenous, which means that a policy change will elicit shifts in expected outcomes and resulting changes in behaviour.

7. NMC rejects both the Friedman 'fooling' hypothesis as well as Lucas' 'monetary surprises' approach to explaining the business cycle. However, NMC does not adopt the view of Keynesians that fluctuations come mostly from the demand side of the economy. NMC is eclectic, allowing for real 'shocks' to both demand and supply. Further, money 'shocks' also have real impacts in the presence of rigid wages and prices. For this reason, monetary policy is effective – money is thus not necessarily neutral.

The NMC can be simplified into a three-equation approach to modelling the economy (see the Chapter Appendix for more technical detail on this). The first equation is a dynamic version of the IS (investment-saving) curve (see Chapter 28) that presents the aggregate demand gap (normally presumed to be positive, that is, with actual output above capacity). The second equation makes inflation a function of the demand gap plus expected inflation. And the third equation presents the central bank's 'reaction function', which describes how the central bank formulates policy to move inflation to its target (following a 'Taylor Rule').

What is important about this framework is that it originally envisaged no role for fiscal policy, instead placing the full burden of adjustment on the the central bank. Further, it assumed that policy would be focused only on inflation, with no concern placed on unemployment. As we will see below, the model was amended after the GFC to create a role for fiscal policy, although its focus remained on inflation rather than on unemployment.

The main idea presented by the NMC is rather simple. If aggregate demand is so high that it creates inflationary expectations, the central bank needs to raise the interest rate; if low aggregate demand induces expectations of deflation, the central bank needs to lower the interest rate. The NMC approach does not rely completely on the interest rate elasticity of spending to stifle aggregate spending. It recognises that increasing interest rates in a boom might not tame speculative euphoria, no more than lowering rates in a slump would spur borrowing and spending. It is thus understood that aggregate spending might not be very interest sensitve.

Instead, what the central bank tries to do is manage expectations. Raising interest rates sends the signal that the central bank is willing to fight inflation, influencing expectations. If firms, households, and markets more generally understand that the central bank will not allow inflation to get out of hand, they will expect inflation to remain under control. In that case, there is no reason to hike prices and wages; with behaviour thus moderated, inflation will remain low. On the other hand, if the problem is deflation, lowering rates signals to markets that the central bank is on guard against downward pressure on prices and wages. This assures workers and firms that deflation will be reversed, raising expectations that prices and wages will resume moderated growth.

The key to control over the economy therefore is to build a consensus about the course of prices and wages; if the central bank can manipulate expectations it will be able to manage the economy. Communication becomes the key. No longer does the central bank try to 'fool' (Monetarism) or 'surprise' (New Classical) markets because its policy works only to the degree that it develops shared expectations.

Hence, the NMC embraces more transparency in monetary policy formation because it believes that policy mostly works by affecting expectations. The central bank signals its intentions long in advance in order to avoid 'fooling' or 'surprises' as it develops consistent expectations in markets. If individuals realise that the central bank is diligently fighting inflation, then they will expect low inflation, and this will cause them to behave in a manner that generates low inflation. For example, slow nominal wage growth would be negotiated in labour contracts on the expectation that inflation will be low over the contract period. Central banks have forsworn 'monetary surprises' (that is, surprise interest rate hikes) in favour of preparing markets for rate changes, and then adopting a method of 'gradualism' in which a series of small rate changes is used to slowly move rates to their ultimate target.

With the onset of the GFC, many central banks, including key banks such as the US Federal Reserve Bank, the European Central Bank and the Bank of England, cut interest rates to historically low levels and engaged in large-scale purchases of government and corporate bonds (the so-called quantitative easing programmes) in an attempt to offset the spending collapse that drove the recession. This was consistent with the NMC view that monetary policy was the most important counter-stabilisation policy tool available. However, when these substantial monetary policy interventions did not stimulate recovery, some NMC proponents argued that monetary policy needed to be complemented by fiscal policy.

Their justification was that when interest rates reach the 'zero bound' (at zero or close to it), the space for further monetary policy interventions becomes limited. In this situation, fiscal policy targeted at promoting higher expectations of inflation may be effective in stimulating spending.

The idea is that in recession, policymakers must create the expectation that a fiscal expansion will spur inflation that will increase future tax revenue to 'pay for' the current deficit. In that case, rational agents will not react negatively to a fiscal deficit. Earlier versions of mainstream theory invoked the concept of Ricardian Equivalence to eschew the use of fiscal deficits. Ricardian Equivalence states that when governments run fiscal deficits, the spending stimulus forthcoming is thwarted by private households increasing their own saving (reducing spending) in order to build up a reserve to pay higher taxes that will be needed to retroactively 'pay for' the deficit. The NMC realises that in some circumstances (where fiscal deficits lead to higher inflationary expectations) this offsetting private behaviour will not be forthcoming.

More importantly, according to NMC, the central bank's low interest rates in the absence of fiscal expansion would actually tighten the government's fiscal stance because it would pay less interest on its bonds. Thus, fiscal policy must be intentionally relaxed to work with monetary policy when it has reached the lower bound at near-zero interest rates. The claim is that after the GFC governments generally failed to do this, which is why monetary policy did not work.

These considerations are believed to relieve government, albeit temporarily, from the dreaded 'budget constraint' (addressed in Chapter 22), but only in the exceptional case where interest rates fall to zero. As the reader will remember, the MMT position is that such thinking is flawed and results from extrapolating the budget constraint of a household to the currency-issuing government.

30.3 Weaknesses of the NMC

The GFC and its aftermath cast unfavourable light on the NMC, and caused its proponents to attempt to address some of its weaknesses. In particular, its treatment of money and financial institutions as well as its relegation of fiscal policy to a subordinate role have been recognised as problematic. In addition, its methodology, which relies on 'aggregating up' from individual behaviour, suffers from fallacies of composition. Finally, its policy recommendations, both before and after the GFC, proved to be rather ineffective.

Proponents of New Keynesian economics and the NMC claim authority because their macroeconomic models are what they call **micro founded**. This just means that they assume people and firms behave as rational, maximising agents with rational expectations and can solve very complex maximisation problems (with respect to consumption and output decisions) about current and future actions. The problem is that the NMC's highly stylised mathematical models, which are overly simplistic because they cannot be 'solved' otherwise, fail very badly when they try to say anything sensible about movements in real world data. At that point, *ad hoc* changes are made to the models (for example, putting lagged variables in to capture real world inertia), which are not indicated by the micro-founded theory. In other words, what eventually emerges as the practice interface of these theories is not based on micro optimisation, and so the claim for authority is negated.

More generally, as argued throughout this textbook, any approach that attempts to explain macroeconomic (aggregate) behaviour by starting out with individual behaviour is exposed to problems of fallacies of composition when it attempts to extrapolate findings to the economy as a whole. Even if individual behaviour can be described as pursuing a rational calculation aimed at utility maximisation over one's lifetime subject to budget constraints, the macroeconomic implications of such behaviour cannot be obtained by simply summing up over a number of individuals. For example, as previous chapters have shown, while spending by individuals could be constrained by income, at the aggregate level it is spending that determines income. Aggregate income is constrained by spending, which in turn is largely determined by expectations (as firms hire the amount of labour they think they need to produce the amount of output they think they can sell at a profit). There are also complex problems of coordination at the aggregate level that are assumed away either by invocation of the 'invisible hand' metaphor or by modelling an economy that has only one 'representative' individual (a typical approach in general equilibrium models) or identical individuals.

NMC economists continue to repeat many of the logical errors of the old neoclassical theorists. For example, proponents of the NMC explained the persistence of the post-GFC recession by arguing that the desire to save outstripped the desire to invest. In their view, this would normally generate an equilibrating fall in the interest rate, which would have reduced saving and increased investment spending. But due to the zero bound being reached, monetary policy could not bring saving and investment into equilibrium.

REMINDER BOX

Students will immediately identify that this argument is based on the flawed loanable funds approach to interest rates that are supposedly determined by the intersection of savings and investment (see Chapter 13). As we demonstrated in Chapter 13, saving equals investment (in the simple model without government or a foreign sector) regardless of the level of the interest rate. Most importantly, saving is a function of income and it is income adjustments rather than interest rate adjustments that bring saving into line with planned investment expenditure. In the expanded model, savings equals investment plus the government deficit plus the current account surplus.

Further, prior to the GFC, the mainstream macroeconomists ignored the financial sector because they believed that it had no relevance to the real sector. In this sense, the New Keynesian approach adopted the classical dichotomy in which money is a veil and is only relevant for determining prices (see Chapter 12). Accordingly, an understanding of the real economy could abstract from the financial sector, with the only concession being the introduction of a central bank following a 'Taylor Rule'.

After the GFC, the NMC economists realised that the lack of any attention to financial markets in their core macroeconomic framework was a major error. A plethora of new academic papers emerged, attempting to integrate banks and financial markets into the New Keynesian model. The revised New Keynesian approach retained the DSGE framework and added elements of the financial sector. Once again this was a case of a practice that heterodox economist David Gordon described in the 1970s as being an *ad hoc* response to anomaly; a characteristic of the neoclassical approach when confronted with major empirical shortcomings.

A full exposition of the technicalities of the DSGE approach is beyond the scope of this textbook. The following introductory statements made to the US House of Representatives Committee on Science and Technology hearing on 20 July 2010 are useful:

> *The dominant macro model has for some time been the Dynamic Stochastic General Equilibrium model, or DSGE, whose name points to some of its outstanding characteristics. "General" indicates that the model includes all markets in the economy. "Equilibrium" points to the assumptions that supply and demand balance out rapidly and unfailingly, and that competition reigns in markets that are undisturbed by shortages, surpluses, or involuntary unemployment. "Dynamic" means that the model looks at the economy over time rather than at an isolated moment. "Stochastic" corresponds to a specific type of manageable randomness built into the model that allows for unexpected events, such as oil shocks or technological changes, but assumes that the model's agents can assign a correct mathematical probability to such events, thereby making them insurable. Events to which one cannot assign a probability, and that are thus truly uncertain, are ruled out.*
>
> *The agents populating DSGE models, functioning as individuals or firms, are endowed with a kind of clairvoyance. Immortal, they see to the end of time and are aware of anything that might possibly ever occur, as well as the likelihood of its occurring; their decisions are always instantaneous yet never in error, and no decision depends on a previous decision or influences a subsequent decision. Also assumed in the core DSGE model is that all agents of the same type – that is, individuals or firms – have identical needs and identical tastes, which, as "optimisers," they pursue with unbounded self-interest and full knowledge of what their wants are. By*

employing what is called the "representative agent" and assigning it these standardised features, the DSGE model excludes from the model economy almost all consequential diversity and uncertainty – characteristics that in many ways make the actual economy what it is.

The DSGE universe makes no distinction between system equilibrium, in which balancing agent-level disequilibrium forces maintains the macroeconomy in equilibrium, and full agent equilibrium, in which every individual in the economy is in equilibrium. In so doing, it assumes away phenomena that are commonplace in the economy: involuntary unemployment and the failure of prices or wages to adjust instantaneously to changes in the relation of supply and demand. These phenomena are seen as exceptional and call for special explanation.

As with all neoclassical general equilibrium models, there is also a problem with trying to introduce money, banks, and the financial system to this stylised framework. For example, the standard DSGE models are not useful for analysing financial crises because debt default is ruled out under the representative agent assumption. This means that there is no need for banks that specialise in assessing creditworthiness. If no one ever defaults then everyone is equally creditworthy and the fundamental activity of banks, what is called 'underwriting', is superfluous. Savers can just lend directly to borrowers or, alternatively, the debts issued by everyone are equally acceptable and always exchange at par against all other debts.

While DSGE modellers want to include money as a medium of exchange to make their theory more relevant to the real world, they have no justification for the existence of banks that would issue it. Given that all debts are risk free, there would be no need for money since any debt could serve the same purpose, you could always buy what you want by directly issuing your own debt. Indeed, debts that pay interest would always trump non-interest paying money. It is ironic that this model is used by central bankers to formulate monetary policy, yet it cannot convincingly justify either the use of money or the inclusion of financial institutions. Attempts to work debt default into the DSGE framework post-GFC add significant complexity and make it largely unworkable.

Economist Willem Buiter (2009), who now works in the financial markets, described New Keynesian and DSGE modeling as "The unfortunate uselessness of most 'state of the art' academic monetary economics". He continued:

Most mainstream macroeconomic theoretical innovations since the 1970s (the New Classical rational expectations revolution ... and the New Keynesian theorising ... have turned out to be self-referential, inward-looking distractions at best. Research tended to be motivated by the internal logic, intellectual sunk capital and esthetic puzzles of established research programmes rather than by a powerful desire to understand how the economy works – let alone how the economy works during times of stress and financial instability. So the economics profession was caught unprepared when the crisis struck .... the Dynamic Stochastic General Equilibrium approach which for a while was the staple of central banks' internal modelling ... excludes everything relevant to the pursuit of financial stability. (Buiter, 2009)

A few years before the global financial collapse, central bankers were congratulating themselves on the success of the NMC approach in not only keeping inflation down, but also stabilising growth and financial markets. In 2004, Benjamin Bernanke, soon to become the Chairman of the US Federal Reserve Bank, declared the arrival of the *Great Moderation* – a new era of stability in which successful policy management by central banks had reduced the risk of run-away inflations or recessions. Central bankers would be able to address macroeconomic problems.

This turned out to be an unfortunate prognosis, as the GFC began just three years later. Alan Greenspan (PBS Newshour, 2008), who was chairman of the US Federal Reserve Bank until 2006, later told the US Congress that the crisis showed that his entire world view, developed over half a century and based on a faith in the efficiency of 'free markets', had been entirely wrong (see Box 32.1). Central bankers had neither understood how the economy worked, nor had they actually produced a new era of stability. In fact, even as Bernanke wrote his paper, the US was living through a period of unprecedented bubbles in real estate markets, commodities markets, and equity markets.

In the aftermath of the crisis, central bankers experimented with historically low interest targets, then turned to unconventional policy such as quantitative easing (QE) and negative interest rates (see Chapter 23), and even discussed policies such as 'helicopter money drops' (in which the central bank would distribute 'free money' to households). The central banks did everything they could think of doing (according to their theoretical positions) to *cause* inflation (and reset expectations) because it remained stubbornly below their targets. While this cast some doubt on the potency of monetary policy, it did not cause a significant change to the new synthesis of macroeconomic theory. New Keynesian economists such as Paul Krugman simply tried to tweak the NMC by arguing that in a 'liquidity trap', monetary policy loses some of its effectiveness (see Chapter 23). While central banks tried QE to lower longer-term interest rates and even negative interest rates by charging interest to banks holding reserves, neither of these had significant effects.

For this reason, as noted above, some NMC economists have begun to move away from the common orthodoxy that fiscal policy is impotent, arguing that at least in some circumstances the 'Ricardian Equivalence' assumption does not hold. In a 'non-Ricardian' situation, an increase of government spending or a reduction of taxes might not be offset by more private sector saving to pay the anticipated higher taxes in the future. In that case, deficit spending can raise demand and thus nominal and even real GDP. While some of the terminology differs from that of MMT, including metaphors such as 'printing money' and 'helicopter drops', at least some advocates of the NMC have come to understand what MMT has long argued.

For example, this is how Woodford (2000: 32, emphasis retained from original) puts it:

> *A subtler question is whether it makes sense to suppose that actual market institutions do not actually impose a constraint … upon governments (whether logically necessary or not), given that we believe that they impose such borrowing limits upon households and firms.* ***The best answer to this question, I believe, is to note that a government that issues debt denominated in its own currency is in a different situation than from that of private borrowers, in that its debt is a promise only to deliver more of its own liabilities.*** *(A Treasury bond is simply a promise to pay dollars at various future dates, but these dollars are simply additional government liabilities, that happen to be non-interest-earning.)* ***There is thus no possible doubt about the government's technical ability to deliver what it has promised.***

Ben Bernanke (2002, emphasis added) reached the same conclusion:

> *Under a fiat (that is, paper) money system, a government (in practice, the central bank in cooperation with other agencies) should always be able to generate increased nominal spending and inflation, even when the short-term nominal interest rate is at zero…* ***The U.S. government has a technology, called a printing press (or, today, its electronic equivalent)*** *that allows it to produce as many U.S. dollars as it wishes at essentially no cost.*

However, the NMC remains a mainstream approach. It still sees taxes and borrowing as the means of financing government spending with money printing as an option to be reserved for extraordinary circumstances, such as a downturn as deep as that of the GFC. Its approach to macroeconomics still relies on aggregating up from individual behaviour, meaning it is subject to various fallacy of composition errors.

It views market forces as equilibrating, even if rigidities, imperfect information, and imperfect competition prevent continuous market clearing. It has trouble introducing money and financial institutions, let alone financial crises, into the analysis in a plausible manner.

Individuals are still presumed to hold rational expectations in a world that is presumed to be actuarially certain. There is no true uncertainty although probabilistic risk exists (see Section 29.5). In all these respects, the world it models bears little resemblance to the world in which we live.

The effectiveness of NMC policy relies critically on expectations management. Both monetary policy and fiscal policy will work only if the central bank can generate a consensus of expectations. For example, lowering inflation requires market participants to expect that inflation will be lower. This lowers actual inflation as firms and workers agree to stop raising wages and prices. When the problem is deflation, it is even more imperative that policy generate expectations of rising prices. Once monetary policy reaches the zero lower bound, it is difficult to do anything *but* to work through expectations management. And as discussed even stimulative fiscal policy will be impotent unless it can produce expectations of inflation due to the notion of Ricardian Equivalence.

Conclusion

Since the early 1980s the central banks have adopted several key principles that were believed to improve expectations management: transparency, telegraphing policy, gradualism, and activism. Transparency means that the central bank will work closely with financial markets, informing them about its policy formation process, providing clear statements about its goals. It telegraphs rate changes far in advance to avoid any surprises (note how different this is from both Monetarism and New Classical views of central bank operations). Gradualism means that the central bank moves rates by small amounts over relatively long periods of time to achieve the total change of rates desired. This also allows markets to adjust gradually to the new interest rate regime. Finally, activism means that the central banks act quickly at the first signs that inflation will move away from target. Ideally, the central bank fights either higher inflation or deflation before it actually appears.

However, in practice, following these principles can prove to be problematic. For example, evidence from the US shows that as the economy recovered from the GFC, the US Federal Reserve Bank began to warn markets that the era of very low interest rates would be coming to an end. This was consistent with the desire for transparency, policy telegraphing, and activism. However, for years after the US Federal Reserve Bank first issued this warning, inflation rates remained below its preferred range. Markets came to expect rate hikes that never happened.

We know from the US Federal Reserve Bank's transcripts that the main reason it raised rates *was because the market expected it to do so, and it did not want to disappoint the market*. In other words, expectations management had gone awry as the US Federal Reserve Bank had built expectations that then forced it to undertake policy that it otherwise might not have undertaken.

Ultimately, market expectations must be linked to reality. After the GFC, the central banks believed that the path to recovery was to generate expectations of inflation. However, as discussed above, even the combination of zero interest rate policy and many trillions of dollars, pounds, euros, and yen spent purchasing through QE policy, the central banks could not induce expectations of positive inflation. This demonstrates that expectations management can be a thin reed on which to hang national economic policy.

References

Bernanke, B. (2002) "Deflation: Making Sure 'It' Doesn't Happen Here", Remarks before the National Economists Club, Washington, DC, November 21.

Bernanke, B.S. (2004) "The Great Moderation", Remarks made at the meeting of the Eastern Economic Association, Washington, DC. February 20.

Buiter, W. (2009) "The Unfortunate Uselessness of Most State of the Art Academic Monetary Economics", *Financial Times*, March. Available at: http://economistsview.typepad.com/economistsview/2009/03/the-unfortunate-uselessness-of-most-state-of-the-art-academic-monetary-economics.html, accessed 25 September 2018.

PBS Newshour (2008) "Greenspan Admits 'Flaw' to Congress, Predicts More Economic Problems", 23 October. Available at: http://www.pbs.org/newshour/bb/business-july-dec08-crisishearing_10-23/, accessed 27 June 2017.

U.S. House of Representatives (2010) "Building a Science of Economics for the Real World", Committee on Science and Technology Subcommittee on Investigations and Oversight, 20 July 2010. Available at: https://www.gpo.gov/fdsys/pkg/CHRG-111hhrg57604/pdf/CHRG-111hhrg57604.pdf, accessed 24 September 2018.

Woodford, M. (2000) "Fiscal Requirements for Price Stability", Princeton University Working Paper, October. Available at: http://www.columbia.edu/~mw2230/jmcb.pdf, accessed 24 September 2018. Published as Woodford, M. (2001) 'Fiscal Requirements For Price Stability', *Journal of Money, Credit and Banking*, 33(3), 669–728.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER 30 APPENDIX: THE NEW MONETARY CONSENSUS MODEL

The NMC is based on an adaptation of the IS-LM model that substitutes an interest rate setting 'Taylor Rule' for the LM curve. The IS curve is modified as an 'output gap' equation, and an equation for inflation is added. The model is easily summarised in three equations:

$$(30.1) \quad Y_t^* = aY_{t-1}^* + bE_t(Y_{t+1}^*) - c[R_t - E_t(p_{t+1})] + \varepsilon_t$$

$$(30.2) \quad p_t = d(Y_t^*) + \alpha_1 p_{t-1} + \alpha_2 E_t(p_{t+1}) + \xi_t; \qquad (\alpha_1 + \alpha_2) = 1$$

$$(30.3) \quad R_t = r^* + E_t(p_{t+1}) + fY_{t-1}^* + h(p_{t-1} - p^*)$$

The coefficients a, b, c, d, f and h in these equations are constants; Y^* is the output gap (the amount by which demand is above or below the full employment level of output); R is the nominal interest rate target; r^* is the 'natural' or equilibrium real interest rate; p is the inflation rate; α represents weights for backward- and forward-looking inflation formation; p^* is the inflation target; and ε and ξ are stochastic shocks.

Equation (30.1) replaces the IS curve, which represents the 'goods' market. The base case is assumed to be a situation of excess demand, driving inflation. If demand were below full employment the gap would be negative, producing deflationary pressures on prices. The gap is a function of last period's gap (so there is persistence in the model), what the gap is expected to be in the next period (forward-looking expectations), the difference between the central bank's nominal interest rate target and expected inflation next period, and a random error term.

Equation (30.2) presents the determination of inflation, which is a function of today's output gap, a weighted sum of last period's inflation rate and expectations of next period's inflation rate and an error term.

The weights sum to one. The model can be made more backward(forward) looking by increasing(reducing) the weight on last period's inflation. Thus, it can become more like Friedman's adaptive expectations model, or more like the rational expectations model of Lucas.

Equation (30.3) is a Taylor Rule, representing monetary policy formation as a reaction to the output gap that drives inflation. The nominal interest rate is set by policymakers depending on expectations of inflation next period, the demand gap last period, and the deviation of last period's inflation from the ultimate inflation target. This then feeds into the IS-like output gap equation based on the assumption that the nominal rate less expected inflation (that is, the real interest rate) influences demand.

PART H

CONTEMPORARY DEBATES

CHAPTER

31 RECENT POLICY DEBATES

Chapter Outline

Learning Objectives

- Understand the competing perspectives about the design of macroeconomic policy to address rising dependency ratios in many developed economies.
- Explain the arguments underpinning the Twin Deficits Hypothesis and their rebuttal.
- Be able to explain the difference between demand- and supply-constrained growth.
- Understand how alternative exchange rate systems have different implications for internal versus external stability.
- Understand how environmental sustainability and full employment can be jointly compatible objectives for a society.

31.1 Introduction

In this chapter we will use our understanding of macroeconomic theory to analyse a number of recent policy debates. On each issue, there is a deep gulf between the arguments made by orthodox economists and the analysis that is informed by heterodoxy – in particular by those who adopt Modern Monetary Theory (MMT). Unfortunately, most of the 'conventional wisdom' held by policymakers as well as by the general public is largely based on orthodox theory. This is because orthodox economists have greater access to the dominant media outlets as well as to politicians. However, as Keynes had warned in 1936, orthodox economics is not only wrong, but it leads to dangerous policy recommendations. MMT provides a useful alternative based on sound theory and policy advice appropriate to the kind of economy in which we live.

31.2 Ageing, Social Security, and the Intergenerational Debate

The material standard of living in a nation ultimately depends on the stock of real goods and services that the nation commands and its ability to access them. Over time, productivity growth provides the means for improving standards of living because it provides a nation with a greater stock of goods and services for a given effort.

When we consider this matter in relation to the broad macroeconomic sectors, government and non-government, the matter becomes more complicated. In a market economy, where goods and services are exchanged for money, the non-government sector's ability to command goods and services depends not only on availability, but also on the sector's capacity to finance purchases.

For the government sector, the ability to command goods and services which are for sale in the currency that it issues depends only on the availability of those resources. That is a fundamental difference which should inform our understanding of the ageing population issue, which is one of the major debates in macroeconomics because of its implications for fiscal policy choices.

The ageing society debate is at the forefront of calls to reduce government deficits. The proposition is that national governments will not be able to afford to maintain the spending necessary to support the growing demands for medical care and pension support as their population ages. At some point, the argument goes, governments will run out of money and other public spending programmes will become heavily compromised.

Some economists support this narrative by advancing arguments about so-called financing gaps, which are attempts to extrapolate future drains on public spending in relation to the projected scale of the economy.

In this section, we will show that these estimates of spending shortfalls are in fact constructed on flawed premises about the fiscal capacity of currency-issuing governments.

Dependency ratios

A motivating factor in the ageing population debate is the fact that dependency ratios are rising in most advanced nations. What is a dependency ratio and why might it matter?

The **total dependency ratio** is normally defined in percentage terms as the population of non-working age divided by the working age population. The demarcation between working and non-working age can vary across nations, but it was typical in most countries for people to be allowed to work after the age of 15 and to retire after reaching 65 years of age.

Accordingly, the working age population (15- to 64-year-olds) is seen to be supporting the young and the old. A rising dependency ratio tells us that over time the share of the population who are productive workers generating national income is falling, whereas the share of the population who are supported by the income is increasing.

If economy-wide productivity were constant over time, then a rising dependency ratio tells us that real average material living standards will decline over time.

Another concept used is the **aged dependency ratio**, which is in percentage terms the number of people above retirement age divided by the number of persons of working age. Similarly, policymakers sometimes refer to the **child dependency ratio**, which is in percentage terms the number of people below the working age divided by the number of persons of working age. The total dependency ratio is the sum of the two. The dependency ratio can be manipulated by policymakers by, for example, increasing the 'retirement age', which reduces the numerator and increases the denominator.

A high child dependency ratio requires higher public investments in infant health, education and childcare, whereas a high aged dependency ratio requires higher public outlays for healthcare and pensions.

With declining fertility rates, the total dependency ratio falls because the proportion of the population in the zero to 15 year group declines. This period in a nation's development has been referred to as the 'demographic dividend' because of the higher proportion of potential workers relative to the number of people that have to be supported. Eventually, as fertility rates continue to fall, the dependency ratio begins to rise as a result of workers transitioning between working age and retirement and improved healthcare outcomes leading to higher life expectancy.

It is argued that the rising aged dependency ratio indicates increased pressure on the public fiscal budget, which is why concern about dependency ratios in the public debate is relatively recent. As a society shifts from one with high fertility to one with low fertility and low mortality, the required public outlays also change in composition with fewer schools and childcare facilities and more aged care homes and pension support being required.

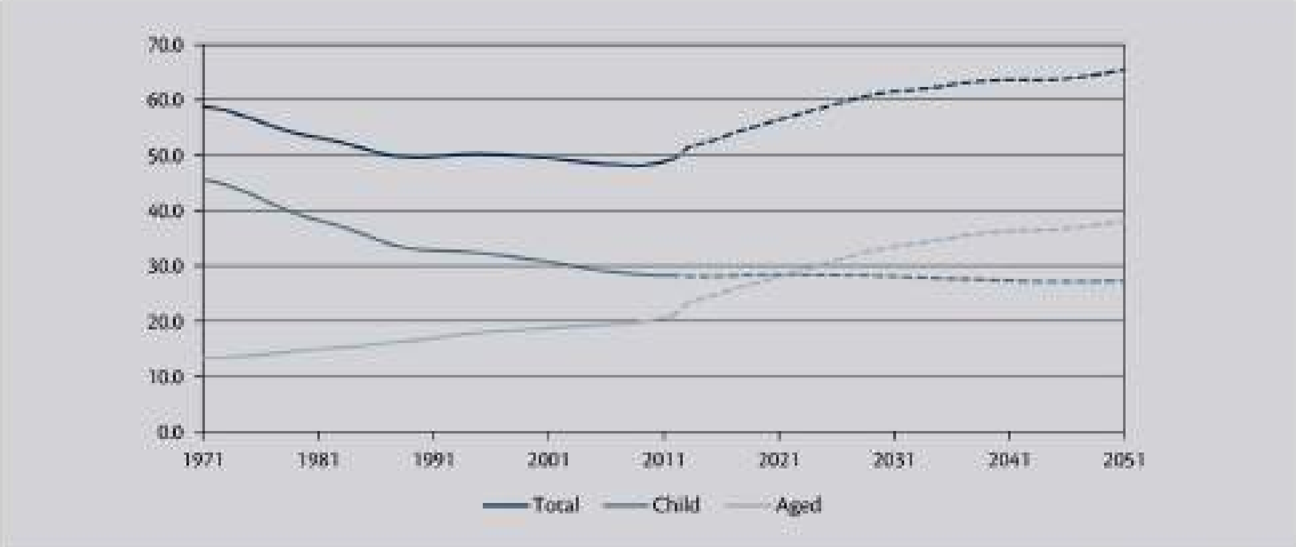
Figure 31.1 shows the total, child and aged dependency ratios for Australia from June 1971 to June 2051. The data from 1971 to 2012 are based on actual population estimates whereas the dotted lines from 2013 to 2051 are based on the Australian Bureau of Statistics Series B demographic projections.

What would be the impact of an increase in the retirement age in Australia from 65 years of age to 70 years? The data show that if the government increased the retirement age from 65 (as it was in 2013) to 70, the dependency ratio in 2051 would fall from a projected 65.4 per cent to 49.9 per cent. In 2013, the total dependency ratio (assuming a retirement age of 65 years) was 51.3 per cent. Thus, policy changes can have rather significant impacts on the dependency ratios.

The standard way of calculating the dependency ratio provides a flawed indication of the relationship between active workers relative to inactive persons, the latter being defined as providing no direct contribution to the production of national income.

The concept of an **effective dependency ratio** has been developed to address these flaws.

Figure 31.1 Total, child and age dependency ratios, actual and projected, 1971 to 2051, Australia



Source: Australian Bureau of Statistics, Demographic data. Authors' calculations.

First, the concept of unpaid work is ignored, which as we saw in Chapter 4, is a standard problem of the conventional national accounting framework. Like all measures that count people in terms of so-called gainful employment, the standard dependency ratio measure ignores major productive activity such as housework and child rearing. The latter omission understates the female contribution to economic growth.

Second, the effective dependency ratio recognises that not everyone of working age, however defined, is actually producing national output and income. There are many people in this age group who are also 'dependent'. For example, full-time students, house parents, the sick or disabled, the hidden unemployed, and those who have taken early retirement fit this description.

The unemployed and the underemployed should also be included in the category of non-productive people of working age, although the statistician counts them as being economically active within the labour force framework. The inclusion of the unemployed and underemployed significantly impacts on the estimated dependency ratio when there is mass unemployment and high rates of underemployment. For example, in August 2013, the Australian labour market data showed that official estimated unemployment was 714,100 and estimated underemployment was 964,310.

Recomputing the dependency ratio (adding these underutilised workers to the numerator and subtracting them from the denominator) produces a total dependency ratio of 69.9 per cent in 2013 compared to the standard estimate of 51.3 per cent. This analysis shows that for Australia (in 2013) the impact of high labour underutilisation on the dependency ratio is more significant than increasing the retirement age by five years.

As we will see, persistent labour underutilisation also exacerbates the implications of a rising dependency ratio because it has adverse impacts on the growth in labour productivity.

These Australian trends are common in most advanced nations where birth rates are falling and people are living longer due to improved nutritional standards and better healthcare.

Do dependency ratios matter?

A dominant view in the public debate is that a rising dependency ratio indicates that more workers will be relying on the state for pension and health support and fewer workers will be contributing to the government's tax base.

The implication of a rising share of the population being economically inactive is that the economically active will have to bear a higher tax burden to support the increased government spending. As time passes, it is argued that the state will enter a fiscal crisis driven by unsustainable deficits and debt obligations.

These claims underpin the dominant rhetoric used by governments, business lobbyists, and economists to justify a preference for the pursuit of fiscal surpluses in the context of ageing populations.

For example, in the US there is constant pressure on the government to privatise the US social security system as a means of keeping it solvent (see Appendix 3 in this chapter for a more detailed investigation of the US social security and Medicare systems). Similarly, in Australia, the so-called intergenerational debate, which began in the mid-1990s and is still a powerful political force, was central to the pursuit of fiscal surpluses by successive federal regimes.

The argument used to support the preference for fiscal surpluses in the context of an ageing population is simple:

- The fiscal deficit-to-GDP ratio cannot be allowed to reach (some projected critical level) because the resulting increases in the public debt-to-GDP ratio would push interest rates up and 'crowd out' productive private investment.
- A higher public debt ratio will impose higher taxation burdens for future generations, which will reduce their future disposable incomes and erode work incentives.
- The retirement age must be lifted so that people will work longer. This will enable them to accumulate more funds to finance their own retirements rather than relying on a state pension. Also, the dependency ratio will be reduced, which raises average living standards.
- Incentives should be introduced to increase the birth rate. For some nations, higher levels of immigration are required to reverse the ageing bias in the population.

What is the veracity of these arguments from the perspective of the macroeconomic framework we have introduced and developed within this textbook?

It is often suggested that fiscal surpluses are equivalent to the accumulation of funds that a private citizen might enjoy as a result of persistent saving. Using this metaphor, accumulated fiscal surpluses are seen as being 'stored away' for the future and thus provide the government with the means to deal with increased public expenditure demands that may accompany the ageing population.

Whether public outlays relating to age-sensitive components of spending rise over time depends on a number of factors, including the nominal reductions in outlays associated with less school-related funding and childcare provision, improved standards of healthcare and the growth in private retirement funds.

However, while the public debate is dominated by considerations of the potential that these 'costs' will inevitably rise, a moment's reflection will assure you that the issue is rather moot.

National government finances can be neither strong nor weak, but merely play a 'scorekeeping' role. We have learned that when government boasts of securing a \$x billion surplus, this is tantamount to saying that the non-government holdings of net financial assets declined by \$x billion over the same period. In other words, private sector wealth was destroyed in order to generate the withdrawal of funds that is accounted for by the fiscal surplus. However, once we appreciate this equivalence, we would conclude that this draining of financial equity introduces a deflationary bias that slows output and employment growth, thereby keeping unemployment unnecessarily high, and forces the non-government sector to rely on increasing debt to sustain consumption.

This idea that accumulated surpluses are 'stored away' and will help government deal with increased future public expenditure demands that may accompany the ageing population lies at the heart of the misconceptions that characterise the intergenerational debate. While it is moot that an ageing population will place disproportionate pressures on government expenditure in the future, it is clear that the concept of pressure is inapplicable because it assumes a binding financial constraint.

The concept of the taxpayer funding government spending is misleading. Taxes are paid by debiting the accounts of the member commercial banks, whereas spending occurs by crediting the same. The notion that 'debited funds' have some further use is not applicable. When taxes are levied, the revenue does not go anywhere. The flow of funds is accounted for, but accounting for a surplus that is merely a discretionary net contraction of private liquidity by government does not change the capacity of government to inject liquidity via net spending at any time it chooses.

The standard government fiscal constraint assumption that deficits lead to future tax burdens because the debt accumulated during the deficit period has to be paid back is also problematic. The government fiscal constraint is not a 'bridge' that spans the generations in some restrictive manner. Each generation is free to select its tax burden via decisions mediated through political processes. Taxing transfers real resources from the private to the public domain.

There is no binding financial constraint on spending by a currency-issuing government, but the government must ensure that its taxation/spending policies are set at the right level to ensure that total spending is sufficient to maintain full employment and is neither inflationary nor deflationary. Thus, the size of the fiscal deficit(surplus) will be market determined by the desired net saving of the non-government sector. The more robust non-government spending is, the lower will be the deficit (which may, under some conditions become a surplus).

This insight puts the idea of sustainability of government finances into a different light. Societies do need to meet the **real** challenges that will be posed by rising dependency ratios.

All societies should aim to provide first class healthcare and viable pension systems, which combine to produce high real standards of living for all citizens. The capacity to achieve those aims depends on the availability of goods and services rather than an erroneous concern about whether the currency-issuing government can afford to purchase them if available. Ensuring the availability of goods and services depends on maintaining high levels of employment and productivity as the dependency ratio rises and adopting environmentally sustainable methods of production and consumption (see Section 31.6).

While lifting labour force participation by older workers is sound, there have to be sufficient jobs available for all workers to make that policy effective. Running fiscal surpluses now and maintaining persistently high levels of unemployment and underemployment is contrary to what is required. Governments should strive to ensure that teenagers are fully occupied in formal schooling, technical training and/or employment. High rates of youth unemployment prevent future workers from acquiring the necessary skills and experience that will underpin a highly productive workforce.

Further, encouraging increased casualisation in the workforce and rising underemployment is not a sensible strategy for the future because it reduces the incentive to invest in one's human capital. Individuals will only lay out resources on education and training if there is an expectation of a return on that investment. The increasing proportion of precarious and low-paid part-time work opportunities in advanced nations is partly due to the excessively tight fiscal positions that governments are choosing to adopt, and reduces the expectation of a net return on skill development.

It is also imperative to invest in public education at all levels to ensure that a highly-skilled workforce is created. Trying to generate unnecessary fiscal surpluses by cutting funding to public education will undermine that essential objective. Public funding of research is also essential to foster increased knowledge accumulation, invention and innovation which will underpin strong rates of productivity growth.

These issues go beyond the realm of economics and government finances. Rather, they focus on political choices. The ability of government to provide necessary goods and services to the non-government sector is not financially constrained. Of particular importance are those goods that the private sector may under-provide. Any attempt to link the two via fiscal policy 'discipline', will not increase per capita GDP growth in the longer term. The reality is that the fiscal drag that accompanies such 'discipline' reduces growth in aggregate demand and private disposable incomes, which can be measured by the forgone output that results.

Running surpluses as a means of stockpiling financial resources for the future does not guarantee that there will be sufficient real resources available in the future. Nations must invest in high productivity strategies to increase the availability of real resources in the future. The distribution of these resources between competing needs will be the outcome of political decisions.

Long-run economic growth that is environmentally sustainable will be the single most important determinant of sustaining real goods and services for the population in the future. The principal determinants of long-term growth include the quality and quantity of capital (which increases productivity and allows for higher incomes to be paid) that workers operate with.

Strong investment underpins capital formation and depends on the amount of real GDP that is devoted to the provision of infrastructure and capital equipment. Public investment is very important in establishing complementary infrastructure upon which private investment can deliver returns. A policy environment that stimulates high levels of real capital formation in both the public and private sectors will engender strong economic growth.

31.3 The Twin Deficits Hypothesis

Introduction

In Chapter 24, we considered the current account in the context of the open economy and the overall balance of payments. The current account is one of the three sectoral balances which, when appropriately defined, sum to zero, as we showed in Chapter 6.

In simple terms, the Twin Deficits Hypothesis states that there is a systematic relationship between two of those sectoral balances, namely the fiscal balance and the current account balance.

A view commonly held by economists is that an increasing current account deficit indicates that a nation is 'living beyond its means', with excessive domestic demand boosting imports and fuelling inflation which undermines the competitiveness of the nation and restricts its exports. They point to the fact that the current account deficit is a measure of the nation's overseas borrowing. Thus, for each quarter that the economy is in external deficit, its stock of foreign liabilities increases. These liabilities can be in the form of debt or equity participation in local firms. Following this logic, an increasing current account deficit indicates that consumption and investment spending in the domestic economy is becoming increasingly dependent on the whims of foreign lenders.

They also argue that there is a strong and direct relationship between the national economy's current account balance and its government fiscal balance and they can become 'twinned', that is, move together dollar for dollar. This is known as the **Twin Deficits Hypothesis**.

Further, the twin deficits argument claims that if the nation has an increasing fiscal deficit, it is increasingly dependent on the foreign purchases of its debt to supplement the purchases by domestic savers.

Thus, according to this narrative, both a rising fiscal deficit and rising current account deficit indicate an increased reliance on foreign debt purchases, which renders the domestic economy vulnerable to unexpected and sudden changes in economic fortunes. Proponents of this viewpoint suggest that the risk of the nation being unable to generate sufficient reserves to pay back the foreign debt rises the longer these deficits occur. Ultimately, the nation would be forced to default on its foreign debt obligations with severe consequences such as the inability to raise capital (private or public) in global markets.

It is argued that the most obvious way to avoid this risky economic environment is to restrict government spending and run restrictive monetary policy (higher interest rates) to encourage increased capital flows even though this strategy reduces the need for them.

The consequences of the imposition of policy austerity are generally agreed to be negative – falling real GDP growth and rising unemployment – which suggests that governments should avoid the twin deficits problem in the first place.

Added to the case against deficits is the claim that they drive interest rates up (through competition for limited loanable funds) and generate inflation (excess demand). High interest rates in turn, are argued to squeeze out productive investment, making the nation less competitive internationally. Competitiveness is also undermined by inflation. This hinders improvements in the trade balance.

The Twin Deficits Hypothesis thus relies on a direct link between fiscal deficits and current account deficits which is considered to lead to a range of undesirable consequences. What is the validity of these claims? As we will see there are several conceptual and empirical difficulties with the view outlined above.

By way of summary:

- The Twin Deficits Hypothesis is based on a particular view about causality, namely that fiscal deficits drive external deficits. In practice, causation may operate in the opposite direction.
- Further, the hypothesis relies on crucial assumptions being made about private domestic spending and saving patterns, situations which have rarely occurred.
- Finally, the crowding-out component of the hypothesis, by which fiscal deficits drive up interest rates and inflation and reduce export competitiveness, is demonstrably false.

In this section, we build on the discussion in Chapter 6 to show that a currency-issuing government that floats its currency has more domestic policy space to work within than is reflected in the Twin Deficits Hypothesis. The government can make use of this space to pursue economic growth and increased living standards, even if this means that the current account deficit increases and the currency depreciates. We also demonstrate that the state of the external sector has very little bearing on what we might consider to be a sustainable fiscal stance.

Finally, it is ironic that many economists who advance the twin deficits relationship as a constraint on public spending also claim that deficit spending is not effective in stimulating national income. The two claims are internally inconsistent.

The link between the deficits

In Chapter 6, we developed the sectoral balances view of the national accounts, which showed us that national income movements ensure that there is a unique relationship between the spending and income balances of the government, external and private domestic sectors.

This framework can be used to demonstrate the accounting basics of the Twin Deficits Hypothesis. The three sectoral balances we derived in Chapter 6 are:

- Private domestic balance $(S - I)$ will be in surplus if the private domestic sector spends less than its income $(S > I)$.
- Fiscal balance $(G - T)$ will be in surplus if the government raises more tax revenue than it spends $(G < T)$.
- Current account balance $(X + FNI - M)$ will be in surplus if exports plus net income inflows are higher than imports and $(X + FNI > M)$.

The sectoral balances equation can be written in the following form (noting we have expressed the fiscal balance as $(T - G)$):

$$(31.1) \quad (S - I) + (T - G) = (X + FNI - M)$$

Equation (31.1) tells us that the external balance $(X + FNI - M)$ equals the sum of the private domestic balance $(S - I)$ plus the fiscal balance $(T - G)$. Other things being equal, an increase in the fiscal deficit will result in a rise (or fall) in the current account deficit (surplus).

We can explain what is going on in the economy by making assumptions about the relative magnitudes of these balances, noting always that the strict equality in Equation (31.1) is always ensured by movements in national income. This is because we assume saving (S), tax revenue (T) and imports (M) are positive functions of income and thus rise when economic activity increases and fall when there is a decline in the level of output and income.

For example, when there is an external deficit ($X + FNI - M < 0$) and a fiscal surplus ($G - T < 0$), there must be a private domestic sector deficit. Consider the following data. If $X = 10$ and $M = 20$ (in the domestic currency), $(X - M) = -10$ (a current account deficit, if we assume $FNI = 0$). If $G = 20$ and $T = 30$, $(T - G) = 10$ (a fiscal surplus). Under these conditions, the private domestic balance, $(S - I)$ will equal $(20 - 30) + (10 - 20) = -20$. This means that the private domestic sector is spending more than it is earning because $I - S = 20$.

Hence, the fiscal drag from the public sector (fiscal surplus) is coinciding with an influx of net savings from the external sector, which finances the private domestic sector deficit spending. Is this situation sustainable?

It all depends on the composition of the private sector spending. If the external deficit reflects private sector borrowing for investment, then the increased productive capacity may generate the returns necessary to sustain foreign debt levels indefinitely at low and stable levels.

If the external deficit reflects a predominance of consumption imports, the capacity of the private domestic sector to sustain the increasing foreign debt liabilities is limited. The private domestic sector can only net save $(S - I) > 0$ when there is a current account deficit ($X - M < 0$), if the public fiscal deficit is large enough to more than offset the current account deficit. Say, $(X - M) = -10$ (as above). Then a balanced fiscal outcome ($G - T = 0$) will force the domestic private sector to spend more than it is earning $(S - I) = -10$. But a government deficit of 25 (for example, $G = 55$ and $T = 30$) will allow the private domestic to save by 15. You can see this by noting from Equation (31.1) that $(S - I) = (55 - 30) + (10 - 20) = 15$.

You can also see that if the private domestic sector is always in balance ($S - I = 0$), then Equation (31.1) can be rewritten as $(T - G) \equiv (X + FNI - M)$, which is a direct relationship between the external and fiscal balances.

Even in this simplified form, we cannot infer causality from the accounting identity. Do higher external deficits ($X < M$) drive higher fiscal deficits ($G > T$) or vice versa? Perhaps the relationship is bidirectional.

The current account position of a nation at any point in time will reflect a range of international factors such as imperfect competition, barriers to entry, economies of scale and general conditions of world trade, all of which are clearly beyond the control of the nation itself.

Consider for example, a world recession which causes a trading economy to experience a current account deficit as its export revenue collapses. Through the operation of automatic stabilisers, the fiscal deficit rises due to declining aggregate demand, real GDP and hence national income and tax revenue. In short, the causality has been reversed. Changes in the external sector beyond a nation's control create corresponding changes in the government's fiscal balance via the automatic stabilisers.

The Twin Deficits Hypothesis imputes a strict causality between the sectoral flows when the gap between private sector savings and investment is zero or stable. Changes in the fiscal deficit translate directly into current account deficit.

This construction is used to justify the claims noted at the outset that a current account deficit represents a nation 'spending more than it is earning' and this imbalance is caused by fiscal deficits which lead to increasing external debt. Accordingly, the risk of foreign financial market retribution via downgrading by international ratings agencies and the like is related to rising fiscal deficits. The cure for a chronic current account deficit then is logically to be found in increased domestic savings, which proponents of the Twin Deficits Hypothesis claim results from fiscal surpluses.

Apart from our discussion in Chapter 9 that exposes the concept of 'government saving' to be erroneous, the other main problem with the Twin Deficits Hypothesis is that its strict causality is not guaranteed. If there is variation in the private domestic balance, then the strict link between the current account and the fiscal balances will not be observed. The evidence from many nations is that the private domestic balance is far from stable over time.

In other words, if for the sake of argument the fiscal outcome was to be balanced over time, and private households and firms started spending more than their income (that is $S - I < 0$), then that private domestic deficit must be funded by the foreign savings, and the current account balance would move into deficit of an equivalent magnitude of GDP as the private domestic deficit.

The underlying economics that would be consistent with this observation might be that increased spending by households on consumption imports and/or spending by firms on investment imports (capital equipment) for a given state of world demand (which sets a nation's export performance) will produce a trade deficit.

Foreign savers will be prepared to fund this deficit because they are willing to accumulate net financial claims in the currency of the deficit nation in return for shipping more goods and services to the nation than it exports abroad.

The opposite would apply if there was a private domestic surplus. In other words, variations in the private domestic balance can drive fluctuations in the external balance quite apart from the state of the fiscal balance.

The Twin Deficits Hypothesis also makes the claim that fiscal deficits drive up domestic interest rates, a response that we considered in Chapter 21. This claim arises from increased competition amongst borrowers for a finite pool of saving which is due to fiscal deficits being financed by debt issue. Higher interest rates then lead to an appreciation in the nation's exchange rate which 'crowds out' exports due to reduced international competitiveness.

We have shown that fiscal deficits do not place upward pressure on interest rates because:

- The official interest rate is set by the central bank.
- The pool of saving grows as national income grows. This observation was a fundamental point of departure for Keynes from the Classical loanable funds doctrine (see Chapter 12). In other words, as seen in Chapter 21, higher government spending (or other spending) generates additional saving via increased national income.
- Banks are always willing to extend credit to customers who they think can meet the terms of the loan. We have learned that loans create deposits, and banks are not constrained by any prior reserve holdings (that is, savings) in their loan-making activities.

BOX 31.1 CASE STUDY: AUSTRALIA

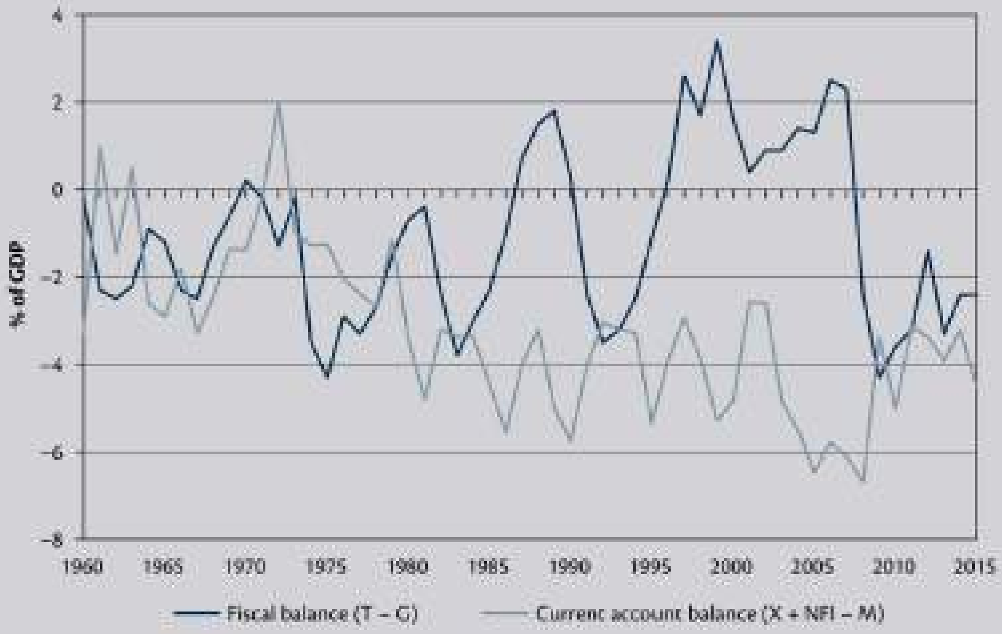
Australia has a very open economy with exports and imports fluctuating around 20 per cent of GDP in real terms. Figure 31.2 shows the movement in the federal fiscal deficit as a percentage of GDP and the current account deficit as a percentage of GDP for Australia for the financial years 1960–61 to 2015–16. From 1983, the Australian dollar was floated.

What would you conclude about the relationship between the current account balance and the fiscal balance from this real world evidence? Does it provide supporting evidence for the Twin Deficits Hypothesis?

Quite apart from causality issues, we would conclude from this data that there has been no systematic relationship between the two balances for Australia over this lengthy time period, which is contrary to the claims of the Twin Deficits Hypothesis.

The current account continued to widen as the fiscal balance went into surplus in the late 1980s and from the mid-1990s. The fact that there is no systematic relationship, however, does not negate the fact that over certain periods, the movement in the external balance could be in response to shifts in the fiscal balance and vice versa. The mediating factor is the behaviour of the private domestic balance.

Figure 31.2 Current account and fiscal balances, Australia 1960–61 to 2015–16, percentage of GDP



Source: Australian Bureau of Statistics, Commonwealth of Australia Fiscal Paper No. 1. Note the fiscal balance is expressed as (T – G) so a deficit is a negative number.

31.4 Balance of Payments Constraints and Currency Crises

Many developing countries have currency sovereignty, which means they can enforce tax liabilities in the currency that the government issues. It doesn't matter if other currencies are also in use in those countries, which is a common situation. For example, the $US will often be in use in a developing country alongside the local currency and be preferred by residents for some of their trading activities. But typically, the residents still must get local currency to pay their taxes. That means the government of issue has the capacity to spend in that currency.

The general principle thus remains as long as there are real resources available for use in a less developed country, the government can purchase them using its currency power. In particular, this concept of real fiscal space extends to the millions of people who are unemployed in less developed countries. Given that there is no market demand for their services, the government in each country can easily purchase these services with the local currency without placing pressure on labour costs in the country. The investment in these programmes is measured by the real resources that they consume relative to not undertaking the initiative. These resources extend to imports and many less developed countries must import food for basic subsistence.

The question then arises: will this investment undermine the current account and introduce inflation as the current account depreciates due to a widening trade deficit?

All open economies are susceptible to balance of payments fluctuations. As we have learned in earlier chapters, these fluctuations are problematic for nations running external deficits under a fixed exchange rate, convertible currency regime (for example, the Bretton Woods scheme) because under these regimes, the government is forced to keep the domestic economy in a depressed state to keep imports down so that the central bank can maintain the parity without losing its foreign currency reserves.

For a flexible exchange rate economy, the exchange rate absorbs some of the adjustment and the balance of payments absorbs the rest. **There is no consistent historical evidence that fiscal deficits create catastrophic exchange rate depreciations in flexible exchange rate countries.**

Generally, an increase in domestic spending will push up import demand. Growth in private capital formation in most less developed countries is likely to be more import-intensive. Encouraging investment-led growth is likely to increase the current account deficit, at least in the short run.

As an alternative, well-targeted government spending can create domestic import competing activity. For example, Job Guarantee (JG) workers could produce goods and services that a nation might normally import, such as processed food products. Moreover, with systems in place to promote the skill development of the labour force, a fully employed economy is likely to attract foreign direct investment under a stable political system. So, while the current account might move into deficit as the economy grows, which means the nation is sacrificing less real resources on exports in return for real imports from abroad, the capital account would move into surplus. The overall net effect is not clear. There are likely to be one-off changes in the exchange rate as the economy adjusts to the higher growth path but these should not be a source of ongoing inflationary pressure.

Finally, a depreciated currency stimulates local employment because imported goods become more expensive and exports become cheaper, but the distributional impacts of these changes are likely to be felt more by the middle- and higher-income groups than the poorer groups because luxury imported goods become more expensive.

It is true that a currency depreciation can be damaging for a nation that is wholly dependent on imported food. Note that this is not a balance of payments constraint as it is normally considered. It is a real resource constraint: insufficient domestic production of food. This can arise from domestic policy choices that are biased against production of food crops, or from the unequal distribution of resources across geographic space and the somewhat arbitrary lines that have been drawn across that space to delineate sovereign states.

In this context, the international institutional framework for dealing with balance of payments difficulties (for example, the IMF and the World Bank) is deficient. It would be improved if a new multilateral institution replaced both the World Bank and the IMF and was charged with ensuring that highly disadvantaged nations can access essential real resources such as food and not be priced out of international markets due to exchange rate fluctuations that arise from trade deficits or adverse financial flows. The simple solution would be that this agency would buy the local currency to ensure flexibility and to ensure that the exchange rate does not price the population, in particular low-income households, out of food. This is a simple solution, and is preferable to forcing nations to run austerity campaigns just to keep their exchange rate higher. The agency should also encourage domestic policy that would increase capacity to produce the output needed by its population, especially food and other essentials.

Currency crises

What about currency crises? Are these inherent to flexible exchange rate regimes, especially those that incur current account deficits?

In the 1990s, there were three major currency crises in the world economy. First, the European currency crisis followed the breakdown of the Berlin Wall. Second came the Mexican currency crisis in 1994, when the peso plunged in value after capital flows, attracted by rising US interest rates, moved against Mexico. Third, there was the Asian debt crisis in 1997.

- What happened in each episode? What were the essential characteristics of the prevailing monetary system at the time of each crisis which precipitated it?
- What lessons can we learn from these relatively recent events?

The European exchange rate mechanism crisis 1992

In 1992, when the German government moved to unify the country after the breakdown of the Berlin Wall, the fiscal expansion required to improve the public infrastructure in the former East Germany, was accompanied by rising interest rates because the Bundesbank (the German central bank) feared inflation. The German mark appreciated significantly because of capital inflow attracted by the higher interest rates and net exports fell.

The problem was that the German mark was the benchmark currency in the European Exchange Rate Mechanism (ERM), against which other Western European nations fixed their currencies. This arrangement had been in place since March 1979. Under such an arrangement, those nations pegging their currencies had to increase their own interest rates in line with increases in rates in the benchmark nation to maintain the fixed parities.

However, the other nations were not undertaking commensurate fiscal expansions to offset the damaging effects of the rising interest rates on their local economy. It became apparent that the commitment to the fixed exchange rate system would mean rising unemployment in these nations and the associated political difficulties.

Players in the foreign exchange markets predicted that eventually the pegging nations would abandon the ERM and let their currencies depreciate against the German mark. This led to the currencies of these nations being sold off in the foreign exchange markets, which immediately forced the nations to reconsider the fixed arrangements.

The impetus to the breakdown of the system was the 'short selling' attack on the British pound by speculator George Soros. In foreign exchange markets, a speculator can contract to sell a currency at some future date for a predetermined price. The contract, of course, means that when the contract comes due, the speculator also has to buy the other currency in the contract. If the contracts are large enough they can thus have a significant impact on the value of the currency and lead other speculators to follow suit. George Soros short sold the British pound against the German mark. His short sales put downward pressure on the pound/mark exchange rate. To maintain the peg, Britain had to buy pounds using its foreign exchange reserves.

Britain left the ERM in September 1992 after the Bank of England had spent more than £6 billion selling foreign currencies in an attempt to maintain the currency (sterling) within the agreed ERM limits while the speculative activity drove its price down. The British government was not prepared to increase its interest rates in line with Germany, a response which they considered would cause a major recession.

In this case, the speculators won, because these governments were intent on pegging their currencies but were not prepared to accept the monetary policy interdependence that came with the decision to peg.

The 1994 Mexican peso crisis

The Mexican peso or 'tequila crisis' was termed "the first financial crisis of the twenty-first century" (Boughton, 2001) by the Managing Director of the IMF, Michael Camdessus, although it was pointed out that, in fact, the Suez was the first crisis and Mexico's problems resembled the factors that precipitated the Suez crisis. The facts are well known although different economists emphasise different causes in their interpretations.

In the 1980s, Mexico endured a debt crisis, which led to misguided policy responses that essentially caused the 1994 currency crisis. The debt crisis was the result of the large foreign-owned commercial banks taking on massive foreign currency-denominated floating interest rate loans. The funds were sourced from the petroleum-exporting nations which had plentiful foreign exchange reserves after the OPEC oil price rises in 1973–74 and later in the 1970s.

In the late 1970s and early 1980s, US interest rates were pushed up to deal with domestic inflation, and the private debt burden for Latin American nations became severe because their interest costs rose rapidly. The US recession in 1981 also damaged primary commodity export markets, a principal source of foreign exchange for the Latin American economies. In 1982, Mexico announced that it could no longer service its external debts and foreign lending to Latin American nations ceased, thwarting the normal refinancing of outstanding loans as they became due. Much of the debt was short term.

The IMF became involved and as its price for the provision of bailout funds, the affected Latin nations had to introduce sweeping free market reforms and severe fiscal austerity measures, consistent with the ideological position of the IMF. Widespread unemployment resulted and poverty rates rose sharply as state-owned industries were privatised, and tariff protection and welfare safety nets cut.

Growth returned in the early 1990s, and capital inflow, particularly from the growing US financial sector, boomed. The free market-oriented Mexican government, under guidance from the IMF, also sought to make the economy as attractive as possible to the financial speculators.

Between 11 November 1991 and 21 December 1994, the Bank of Mexico managed a fixed peso parity against the US dollar within a so-called 'slippage regime' where the rate was allowed to vary within certain daily bands. The upshot was that the Mexican government stood ready to convert the peso into US dollars (and vice versa) at a fixed rate, which meant it always had to have sufficient reserves of US dollars to guarantee convertibility. Reliable convertibility was thought to be essential to establish the monetary credibility of the Mexican government and instil confidence into the international financial markets. The reality was that all the risk had been shifted from the foreign speculators to the Mexican government; a risk that ultimately the Mexican economy was unable to bear.

Mexican domestic growth increased the current account deficit, which was met by massive capital inflows. Under pressure from the IMF, the government was running a small fiscal surplus, but it was forced to keep issuing government debt to foreign creditors to provide financial instruments that would attract increasing capital inflow. The domestic growth also increased inflation as productive capacity was stretched.

These developments meant that the peso should have depreciated but the Mexican government, under pressure from the IMF and the US government, maintained the peg even as speculators started to sell off the peso in favour of the US dollar. The Bank of Mexico's stock of foreign exchange reserves (so-called 'hard' currency) started to run out and this led to further speculative attacks.

There was some political turmoil, which didn't help the declining confidence in the Mexican economy.

The short-sighted commitment to fixed parity by the Mexican government led them to further compound the crisis when they acceded to demands from large foreign investors (particularly Wall Street banks) to significantly increase the issuance of so-called Tesobono bonds, which were US dollar-denominated government debt instruments that insured the holder against any foreign exchange risk.

The foreign currency-denominated bonds were increasingly substituted for peso-denominated debt to keep the foreign investors happy. The foreign (exchange rate) risk insurance explicit in the Tesobonos lowered the price the government had to pay on this debt but also dramatically increased its exposure to peso depreciation.

Despite the tension within the government and between the government and central bank, the fixed parity position was maintained even though it was obvious that the government could no longer support the currency. However, the IMF continued to argue that the Mexican policy settings were sound. As late as its Executive Board Meeting of 30 November 1994, the IMF discussed the introduction of a new 'short-term financing facility' for eligible nations to assist in the alleviation of short-term balance of payments pressure that might destabilise their currencies.

However, less than a month later, the speculative outflow of pesos became too great and on 22 December 1994, the Mexican government floated the peso. On 19 December 1994, the peso was trading at 3.4662 per US dollar. Eight days later, the parity had risen to 5.7625 per US dollar, which represented a depreciation in the peso of 66.2 per cent.

The decision by the Mexican government to float the peso was forced on it because it had run out of the foreign reserves necessary for the central bank to maintain the peg against the US dollar, due to speculators selling the peso and driving its price down in world markets.

The short-term consequences of the depreciation were severe. This severity was linked, in part, to the government's tardiness in making decisions. The events occurred in a presidential election year and both the instability associated with a bitterly fought campaign (including the assassination of one of the leading candidates, Luis Donaldo Colosio, in March 1994) and the reluctance of the incumbent to bear the stigma of allowing a devaluation, meant that the peg remained in place for too long despite massive outflows of funds.

The peg signified a sort of status for the Mexican government, which had instilled a sense of confidence in its economy by Mexico becoming a member of the Organisation for Economic Cooperation and Development (OECD) and entering the NAFTA in early 1994.

The government had also liberalised credit and privatised the banking system, which exposed the nation to rapid capital inflow without a commensurate increase in the ability of its financial institutions to handle the risks involved in international finance. Once Mexico was forced to float, the response of the international markets was extreme. The investors (both foreign and Mexican), who had previously held out Mexico as the exemplar for Latin America to follow, sold off pesos in astonishing proportions in a space of a few days (20 to 22 December 1994).

Given the desire to maintain the peg, the government left itself with two undesirable options as the capital outflow accelerated and the central bank foreign reserves rapidly declined.

First, they could have hiked interest rates to encourage investors to leave their funds in Mexico. However, the required interest rate increases would have been so large that they would have plunged the economy into a major recession, which might have reduced confidence further and added more pressure to the declining peso.

Second, they could have broadened the bands in which they allowed the peg to crawl, which would have improved the current account a little and perhaps offset some of the mania that was creeping into foreign exchange markets about the likely depreciation of the peso. But the consequences of that option would be to inflate the debt-servicing payments on foreign debt held by Mexican companies and to leave the Government with the inevitable consequence of insolvency.

Eventually, a combination of IMF and US government assistance stabilised the financial system and placated the international investment community, which belatedly realised that the economic fundamentals of Mexico were unchanged and had not justified the massive overreaction.

In summary, the Mexican peso crisis teaches us some important macroeconomic policy lessons:

- While a floating exchange rate may expose an economy to imported inflation in times of depreciation, the advantages of being able to stabilise domestic output and employment are significant. A nation which pegs its currency loses control of monetary policy, and forces fiscal policy to play a passive role that becomes destructive when the currency depreciates significantly.
- While a currency issuing government can issue public debt, it is imperative that any liabilities it does issue are denominated in its own currency and that it assumes no foreign exchange risk by way of indexing or insurance arrangements.

The South East Asian debt crisis 1997

The last major currency crisis of the 1990s began in 1997 in the South East Asian nations of Thailand, Malaysia, the Philippines, Singapore, Indonesia, and later spread to the industrialised East Asian nation of South Korea.

In the two decades leading up to the crisis, the South East Asian nations had attracted large capital inflows and their economies had grown considerably. The period of rapid growth, which started in the late 1980s, was accompanied by high private saving ratios and strong investment. Further, inflation was low and the governments were largely running fiscal surpluses.

The South East Asian economies were considered by the IMF and the World Bank to be models of sustainable development. The expression 'The Asian Miracle' was used to describe the rapid growth and rising living standards, particularly in the so-called four 'Asian Tiger' economies of Hong Kong, Taiwan, Singapore and South Korea. The multilateral organisations mistakenly believed that the rapid growth was the product of fiscal rectitude and free market dynamics, which they thought diverted resources to their highest value use and allowed these nations to be internationally competitive.

However, the reality was different. The Asian nations built their growth strategy, which began in the late 1960s (with Japan), on a mix of industrialisation, mercantilism and strong state-imposed industrial policies.

In nations such as South Korea, the state played a major role in the development process and defied the advice offered by free market economists at the IMF and the World Bank with respect to their development strategy. The international organisations considered that trade-led growth would only come if a nation exploited its comparative advantage. However, the South Korean government selected and supported several key sectors to be their growth engines, despite none of them having any relative resource advantage (for example, chemicals). The textiles sector in Korea had indicated that a chemicals industry would support its own development.

The governments in fact interfered with the 'market' in many ways. They provided credit at below-market prices to targeted sectors. Substantial tax breaks were given to firms to increase profits and investment. Protection was provided to local firms against import competition. The state invested heavily in public research and development and shared the results with industry.

By the early 1990s, capital resources were shifting from the Tigers to China and India attracted by cheaper labour resources. The growth of China and India challenged the export supremacy of other Asian nations such as Malaysia, Taiwan, South Korea, Singapore, and Thailand, which had led the Asian growth phase in the late 1980s. The shifting investment and rising export strength of China reduced growth rates in the Tiger nations.

Several other shocks occurred in this period, which undermined the miracle. First, the Chinese renminbi and the Japanese yen were devalued. Second, the US Federal Reserve increased interest rates, which pushed up the value of the US dollar and placed strain on the currencies which were pegged to it. Third, the fall in the Tigers' export earnings was exacerbated by the large global fall in semiconductor prices (a 36 per cent decline between 1993 and 1999).

The growth phase had also been accompanied by a boom in real estate prices, which was fuelled by significant short-term foreign currency loans, increasing the risk exposure of a private sector that was reliant on high export incomes to service the debts.

The crisis proper began in Thailand in July 1997. The Thai baht was pegged to the US dollar, a practice that was common among the Asian economies. Its real estate sector had pushed the nation's foreign debt beyond sustainable limits, and speculative capital outflows, motivated by the fear of losses if the currency fell in value, put pressure on the exchange rate. In the face of these pressures, the central bank was unable to maintain the peg as it ran short of the required foreign currency reserves. Once the government floated the baht on 2 July 1997, its value fell by more than 50 per cent as international investors dumped it on the foreign exchange market and created a massive excess supply.

The collapse of the currency effectively rendered the nation bankrupt, given the large volumes of foreign currency-denominated debt held by the private sector. Since much of their earnings would be in devalued domestic currency, they could not exchange it for enough foreign currency to service the debt. There was a significant fall in the local stock market and several major financial institutions were bankrupted.

The crisis exposed the dangers of maintaining currency pegs which required central banks to have sufficient foreign currency reserves to maintain the agreed parities. This made all currencies in the region susceptible to speculative attacks.

While the structure of the Thai economy was very different to the Tigers', speculators considered that all regional currencies were in similar danger. This belief became a self-fulfilling prophecy and by August 1997, speculative attacks on the currencies of Indonesia, Malaysia and the Philippines led to declines in their exchange rates. The crisis spread in September to Hong Kong, Singapore, and Taiwan, and in November 1997, the capital outflow from South Korea forced it to devalue. The banks that had extended short-term loans to these nations refused to roll over the debt and an instant credit crunch was created.

It is clear that the stronger advanced nations such as the US, Japan and the EU could have intervened and facilitated enough liquidity to stop the capital outflow which had resulted from the panic. Not only could their central banks have provided credit lines to the central banks in the Asian nations, but the advanced governments could have brokered roll-over arrangements with the private banks to stop the panic. Instead, the main response from the advanced nations came through the IMF, which intervened first in Thailand in July 1997. Immediately before the crisis, the IMF considered the Asian economies – both advanced (South Korea) and less developed (for example, Indonesia and Thailand) – to be growing strongly on the back of extensive deregulation of their financial systems. It believed that further 'free market' reforms would be beneficial. While there was some recognition that capital inflow was very strong and perhaps volatile, the IMF failed to correctly assess the vulnerability that their policy prescriptions (liberalisation and so on) had created.

By the end of 1997, the IMF was harshly criticising the Asian governments that earlier in the year it had been praising.

It is now accepted that a series of policy blunders in the IMF response deepened and spread the crisis. In return for bailout funds, the IMF insisted that the nations under speculative attack in the currency markets introduce sharp increases in interest rates and substantial fiscal contraction. The IMF applied their 'one size fits all' standard response for when fiscal deficits were significant and inflation accelerating. This approach is questionable at the best of times, but certainly it was inapplicable to the Asian economies, which were running fiscal surpluses and had stable inflation rates.

In its most simple form, the crisis was the result of excessive financial liberalisation that promoted massive capital inflow (and commensurate liabilities) between 1993 and 1996. The liabilities tended to be short term but the funds were used for long-term investments (for example, real estate purchases).

When export growth slowed, the capital inflows started reversing very quickly, but the IMF demands ensured that the crisis moved out of the foreign exchange markets and became a full-blown economic recession. As capital outflows accelerated with the worsening economic conditions, the IMF insisted that interest rates be pushed up further and fiscal contraction deepened.

As part of the Indonesian bailout plan, the IMF forced the government to close 16 insolvent banks, claiming that this would restore confidence in the remaining banks. The result was the opposite and the panicked withdrawals of funds undermined the solvency of many of the private banks. The Indonesian central bank injected funds (equivalent to five per cent of GDP) into these banks to save them, which had the effect of exacerbating the collapsing rupiah and was at odds with the IMFs insistence that interest rates had to rise sharply.

In summary, the Asian financial crisis was the result of a lack of regulation of capital flows combined with the currency pegs. In the case of the latter, these were interpreted by financial markets as the government insuring them against foreign exchange risk and so there was a lack of private foreign exchange hedging of the borrowing. Once the currencies collapsed and were floated, these unhedged positions quickly led to bankruptcy.

31.5 Fixed versus Flexible rates: Optimal Currency Areas, the Bancor, or Floating Rates?

Introduction

Throughout this text we have discussed the link between a sovereign currency and domestic policy space. Most mainstream economists either do not recognise the importance of this link, or they reject it. Indeed, some argue that it would be better if nations did not issue their own sovereign currency because it tempts them to use 'money printing' to finance excessive deficits that cause inflation.

There are several ways that the link between the currency and policy can be broken: adopt a foreign currency for domestic use ('dollarisation' or currency board arrangements are examples); adopt a shared currency (the European Monetary Union created a new shared currency to be used by all members); or adopt the Optimal Currency Area (OCA) proposal developed by Robert Mundell (1961). In important ways, the OCA approach provides a coherent integration of the first two alternatives.

Optimal currency areas

Mundell argued that if money is first and foremost a medium of exchange, it makes little sense for the use of a currency to be contiguous with the nation state. After all, national borders are largely politically determined. It is certainly possible that the 'economic borders' of a highly integrated economy would not coincide with the political borders of the nation states. For example, perhaps the eastern US and Canada might be more economically integrated than the eastern and western areas of each individual country are with each other. Hence, it might make sense to have a 'western' dollar for the western regions of each country and an 'eastern' dollar for the eastern regions of each country, rather than a US dollar and a Canadian dollar. Another example might be the cases of Estonia and the Czech Republic, each of which has regions that are highly integrated with the German manufacturing sector; so maybe they should share the same currency? (Of course, Germany and Estonia do share the euro, while the Czech Republic has its own currency; it could be argued that Estonia and the Czech Republic are far more integrated with Germany than is Greece, which also uses the euro.)

An optimal currency area should be one in which capital and labour are mobile and wages and prices are flexible within the region. This allows resources to move where they are needed, and wages and prices to adjust to maintain equilibrium. Within the currency area, the exchange rate is obviously fixed because the entire region uses the same currency. It is claimed that flexible capital and labour markets can then operate to equalise supply and demand (since it is not possible for exchange rates to adjust to bear some of the burden of maintaining market equilibrium). It is also important that all areas within the region are subject to the same business cycles, so that the monetary and fiscal policies chosen in reaction to the cycle are appropriate across the region. Flexible exchange rates with other currencies (those outside the OCA) maintain external equilibrium; that is, current accounts are supposed to be balanced by flexible exchange rates against regions (and countries) outside the OCA.

In addition to constraining profligate governments, the OCA is supposed to reduce the transaction costs involved in exchanging currencies (fees that are charged, plus the costs of dealing with uncertainty over movements of the exchange rate). No one has to carry a calculator to try to convert prices across exchange rates within the currency area. Businesses do not have to protect themselves from unfavourable movements of exchange rates (without the OCA, their costs might be paid in one currency, but their revenues would be in another if they were exporters). Moving to a single currency reduces costs and uncertainties, increasing beneficial trade across national borders. The region as a whole benefits.

The OCA proposal was extended to a call for the division of the globe into a small number of huge regions, each to use a single currency. For example, all of the Americas might adopt the US dollar; the euro region would be extended all over Europe and perhaps into Africa; and either the Chinese RMB or the Japanese yen would be adopted across Asia. Within each region exchange rates would be fixed, but each of the major regional currencies would be flexible with one another. In this way, exchange rate adjustments would balance payments across regions, and flexible markets within regions would maintain full employment. With individual nations abandoning their currencies, markets would discipline governments to pursue balanced budgets and low inflation.

Mundell's theory was often invoked during the arguments for the creation of the European Monetary Union (EMU). Even if the nations that were to join did not precisely fit the preconditions enumerated above, it was argued that by abandoning their currencies and adopting the euro they might *become* an optimal currency area. We have discussed in detail the outcome of the creation of the EMU. Most analysts today have (belatedly) come to realise that the fatal flaw in Mundell's proposal was his failure to recognise that national currency sovereignty is the critical determinant of fiscal and monetary policy space. Abandoning their former currencies had devastating consequences for most of the members of the EMU.

If it is accepted that nation states need their own sovereign currencies to support independent fiscal and monetary policy, that still leaves up in the air the proper international monetary system. In this text we have taken the position that flexible exchange rates preserve domestic policy space. In the remainder of this section we examine alternative international payments systems, beginning with the development of the system adopted at the end of the Second World War.

The demise of the gold standard: the Great Depression and the Second World War

Before the Great Depression, many of the major capitalist countries had adopted a gold standard, promising to exchange their currency for gold at fixed exchange rates. However, during the Depression, they abandoned gold convertibility in an effort to free fiscal and monetary policy in order to deal with the calamity. The Second World War, which followed the Depression, pitted nation against nation, making normal trading relations impossible. As the end of the war neared, the Allies began to plan for a post-war world and they discussed the kind of international currency regime that would be put in place.

It is important to understand that before the war, the British pound had played a large role in international payments. Although countries were on the gold standard, payments were often made in pounds (at the fixed parity to gold). But Britain's status had fallen while America's status had risen during the war. At an international meeting in 1944 in Bretton Woods, New Hampshire (US), leaders from many of the nations of the world discussed the post-war system. J.M. Keynes represented the UK while Harry Dexter White represented the US in the negotiations. Each brought a plan to the meeting; Keynes argued for a 'Bancor' plan, while White argued for what would come to be called the Bretton Woods System. In addition to setting up the international exchange rate system, the outcome of the conference included the creation of the International Monetary Fund (IMF) and the World Bank.

The Bretton Woods System (BW) largely conformed to the US's preferences. The US would fix its dollar to gold while other nations would fix their exchange rates to the dollar. International payments would be made largely in dollars. The IMF and World Bank would help to promote development by providing more flexibility to the system than had been possible under the pre-war gold standard. Nations that needed dollars to protect their exchange rates against current account deficits could turn to the IMF and World Bank (in addition, the US provided dollars directly through its Marshall Plan).[1] However, this system was still more rigid than the one Keynes had brought to the table, and lasted only a quarter of a century before being abandoned in the early 1970s. It is interesting to look at the features of Keynes' proposal, which might have been more successful. In recent years, critics of flexible exchange rates and of the dominance of the US dollar in global financial markets have renewed calls to return to Keynes' proposal: to create a second, better, Bretton Woods system.

Keynes' Bancor plan and the end of Bretton Woods

The discussion here follows the work of Sardoni and Wray (2007). Keynes called for the creation of an International Clearing Union (ICU) based on a unit of account called the Bancor. The Bancor would be fixed in value relative to gold and then the currencies of all countries participating in the ICU would be fixed relative to the Bancor. The Bancor would be used only for clearing purposes among countries; countries could buy Bancor balances from the ICU using gold, but Bancors could not be redeemed for gold, ensuring there could be no run on Bancor.

At the outset, the quantity of Bancor reserves would be distributed among countries based on previous levels of international trade. Countries running trade surpluses would accumulate additional reserves, while deficit countries would lose reserves. The ICU would provide overdraft facilities to those countries that exhausted their reserves. Reserves could not leave the system, and the ICU could always expand the supply of Bancor reserves by making advances to deficit countries. In addition, surplus countries could use Bancor reserves to make loans to, investments in, or unilateral grants to deficit countries.

Keynes called for a charge on excessive overdrafts and on excessive reserve balances of one or two percentage points in order to encourage balanced trade. Other possible actions to be taken in the case of deficit countries included currency devaluation, capital controls, seizure of gold reserves, and domestic policy. Actions to be taken against surplus countries included requiring the expansion of domestic demand, appreciation of the currency, reduction of tariffs and other trade barriers, and encouragement of international development loans (Keynes 1980: 462–3). Finally, the ICU could use its power to encourage economic development through the use of overdrafts for relief work, for development of buffer stocks of commodities, for the establishment of an International Investment Corporation, and to help stabilise prices (Keynes 1980: 190).

The Bancor Plan was never adopted. The Bretton Woods system worked so long as the flow of dollars from the US to the rest of the world was just sufficient to meet the world's demand for dollar assets. By the late 1960s, the dollar was under pressure, the quantity of dollar claims around the world greatly exceeded the US gold reserves and there was fear that holders would exhaust these reserves if only a relatively small portion of dollars were submitted for redemption. Indeed, any hint of significant devaluation of the dollar would generate a run. Fearing this could happen, the system was abandoned by President Nixon in the early 1970s in order to protect the US gold reserves. The US floated its currency and most large nations followed suit.

Note that the weaknesses of the BW system in comparison to Keynes' Bancor plan included:

1. The US dollar was used as the international reserve currency; in Keynes' plan, the reserve currency would be the Bancor; an international currency rather than the currency of any individual nation.
2. The US dollar was convertible to gold, which made runs possible (the Bancor was not convertible).
3. The 'reflux' method of sending international currency reserves to deficit nations was far weaker under the BW system than under the Bancor plan.
4. No penalties were put on surplus nations under the BW system, while the Bancor plan would force nations that accumulated excessive reserves to 'use or lose' them.

When inevitable pressures on the dollar led to the collapse of the BW system, the major countries joined the US in moving to floating exchange rates. The move was supported by some neoclassical economists, most notably by Milton Friedman. The claim was that a flexible system would operate much like the fabled 'specie flow' mechanism purported to have quickly rectified trade imbalances during the gold standard era. The orthodox claim was that, in a flexible exchange rates regime, trade imbalances would cause currencies to adjust (surpluses would appreciate the currency, deficits would depreciate it), automatically returning global trade to equilibrium.

After nearly a half-century of experience since the abandonment of Bretton Woods, this claim is thoroughly discredited. To be sure, many nations have not adopted freely floating exchange rates, but interventions are not sufficiently large to explain the persistence of trade imbalances. In addition to long-term trade imbalances, the world has experienced a great deal of exchange rate instability (especially among developing nations) that is far larger than would be expected simply because of trade imbalances. Thus, the move to flexible exchange rates has not resulted in balanced trade. Nations like the US have run large, and rising, current account deficits for years while nations like Japan ran large current account surpluses for decades without setting off equilibrium-seeking exchange rate adjustments. This experience has led some to call for a new BW system, or for a reconsideration of Keynes' Bancor plan.

One difficulty with Keynes' approach to exchange rates and reformation of the international monetary system along the lines of the Bancor plan is that like many mainstream analyses, it concentrates on trade imbalances and the current account, while capital movements receive too little attention. To be fair, Keynes did not foresee, nor would he have supported, 'free' capital flows. His proposal presumed that finance would remain national. Note also that the Keynes proposal preserved the option of imposing austerity on nations with persistent trade deficits, recognising that exchange rate adjustments might not work. As a consequence, money tends to be seen as a medium of exchange. In his work on the ICU, Keynes began by noting that his goal was to design an international currency system in which the currency exchange will be made to operate as if countries were "trading goods against goods" (Keynes 1980: 18). The operation of the ICU would be designed to ensure that Bancor reserves would not be lost to idle hoards; rather, the reserves of one country would form the basis of overdrafts of another, thereby encouraging trade. While Keynes' proposal would penalise nations that accumulate international reserves, it still appears to be based on the view that money (mostly) circulates goods.

In the real world today, currencies are not used solely for current account transactions, but also in capital account transactions. Therefore, unless capital is completely immobile, there is no reason for exchange rate adjustments to eliminate current account imbalances. Indeed, with the growth of global finance that easily evades national restrictions, the vast majority of international transactions do not directly relate to current accounts.

If capital controls are neither politically nor technically feasible, then designing an exchange rate system based on the belief that the international monetary system should operate as if goods trade against goods is unlikely to function well. Further, it is hard to imagine countries and groups of countries with interests as diverse as those of the US, China, and Euroland coming together to create anything like an international clearing union with the power to penalise trade surplus countries such as Japan, Germany and China, let alone to impose austerity on the world's biggest trade deficit country, the US.

The Keynes plan hoped to enhance external stability by fixing exchange rates and adopting several procedures to reduce trade imbalances. However, for reasons we discuss below, this comes at a great cost because the ability of a nation on a fixed exchange rate to use domestic policy to achieve internal stability is reduced. If the enhanced external stability does not itself also generate internal stability, there is a trade-off of external stability for internal instability.

An alternative (MMT) approach to international money: floating rates and sovereign currency

As we argued in the previous section, reformist projects for the international monetary system that concentrate on trade and current account imbalances are an inadequate response to the present situation, in which capital movements play a central role. Here we outline an alternative approach to international money regimes, which is also derived from Keynes. This alternative approach is relevant in the present context because it implies the adoption of a regime of floating exchange rates by countries.

A nation like the US (as well as countries like Japan, Britain, the European nations before they adopted the euro, and Argentina after it abandoned the currency board) creates a currency for domestic use and ensures its use primarily by demanding payment of taxes in that currency, although some also adopt legal tender laws. The state (including the treasury and the central bank, which acts as an agent of government) issues and spends the monetary base (cash and reserves at the central bank) without any promise to convert the monetary base to any other currency, or to gold or any other commodity, at any fixed exchange rate. The ability of a national state to behave in this way with respect to its currency and to maintain its fiscal independence is what we mean by our reference to sovereignty here.

The sovereign government's ability to make payments is neither revenue constrained nor reserve constrained. Further, the interest rate paid on sovereign securities is not subject to normal 'market forces'. Since short-term government debt is essentially equivalent to interest paying reserves, and thus a close substitute to lending in the overnight interbank market, the overnight rate set by the central bank will govern the short-term government 'borrowing' rate.[3] This indicates that a sovereign nation can choose short-term interest rates on government debt as low (or high) as it wants. Whether the base rate will be zero or one hundred is a monetary policy matter, not subject to market determination.

A non-sovereign government is in an entirely different situation. In a 'dollarised' nation, the government must be able to obtain dollars. While it might be able to issue its own IOUs denominated in (say) US dollars, to maintain parity against the dollar it must stand ready to convert its IOUs to actual US dollars. Hence, it uses taxes and issues IOUs to obtain US dollar reserves in anticipation of spending; unlike the case of a sovereign nation, this government must have deposits of dollars before it can spend. Further, unlike the sovereign nation, the non-sovereign government promises to deliver third-party IOUs (US dollars) to service its own debt (while the US and other sovereign nations promise only to deliver their own IOUs). Because of this, the interest rate on the non-sovereign government's dollar liabilities is not independently set. Since it is effectively borrowing dollars, the rate a dollarised nation pays is determined by three factors. First there is the base rate on dollars set by the monetary policy of the US (the issuer of the dollar). Second, there is the market's assessment of the non-sovereign government's creditworthiness, which may be determined by a large number of factors. These two considerations determine the lowest interest rate the market will tolerate. Third, the interest rate is also constrained by the nation's need to keep its exchange rate fixed or pegged to another currency. Thus, a non-sovereign government, as user (not issuer) of a currency, cannot independently set the domestic interest rate.

From this it follows that the capacity for a state to set the interest rate in a world of high capital mobility is contingent on its adoption of a floating exchange rate regime. A regime of floating exchange rates is important for developing countries in particular. They are rightly concerned with the financial and exchange rate crises suffered not long ago by Asian (1990s) and Latin American (1980s) nations (discussed earlier in this chapter) that were triggered by large external debts, declining foreign currency reserves, and market expectations that exchange rate pegs could not be held.

By contrast, a nation that adopts its own floating rate currency can always afford to put unemployed domestic resources to work. Its government will issue liabilities denominated in its own currency, and will service its debt in its own currency. Whether its debt is held internally or externally, it faces no insolvency risk. This does not mean that the nation can necessarily ignore its trade balance or movements of its exchange rate, but it does mean that it can put domestic employment and growth at the top of its policy agenda.

A sovereign nation is able to use domestic policy to achieve domestic or internal stability. This comes at the cost of possibly greater external instability. A floating exchange rate will not necessarily move trade toward balance, as discussed above. However, it must be remembered that from the macro perspective, imports are a benefit while exports are a cost. Hence, a trade deficit means net benefits. This is usually neglected in discussions of trade balances because of the presumed impacts on domestic employment. But, so long as the nation's domestic policy is geared toward stability, it can achieve full employment even in the presence of a trade deficit.

This in turn requires currency sovereignty, which necessitates a floating exchange rate. It is possible that a trade deficit can exert downward pressure on the exchange rate, which can generate some 'pass through' impacts on domestic inflation. If desired, domestic policy can turn to inflation fighting, including the conventional method of adopting tighter fiscal policy to attenuate inflation pressures. However, using its sovereign currency the nation can instead adopt a job guarantee that automatically ensures full employment while also helping to stabilise wages and prices.

A nation that adopts a fixed exchange rate must hope that the conditions that generate external stability will also happen to coincide with those that permit internal stability. By contrast, the nation that floats can enjoy the net benefits of a trade deficit, improved real terms of trade (a trade deficit means that the 'real' cost of imports in terms of exports is lower), and domestic full employment; all of which might be somewhat offset by the possible costs of currency depreciation and higher prices. The nation that fixes the exchange rate may not be able to 'afford' a trade deficit (because of exchange rate pressures), and will probably have to use domestic unemployment as the means to maintain its peg. For these reasons, a flexible exchange rate preserves 'policy space' for independent policy formation.

The euro and optimal currency areas

The European experience leading to the creation of a common currency area is another important example of the detrimental consequences for national states of giving up their sovereignty through renouncing flexible exchange rates. European countries adopted a common currency area and created a single central bank in their quest for stable exchange rates within Europe. There is nothing inherently wrong with monetary integration that leads to a fixed exchange rate within a union; indeed, the US can be thought of as a currency union with fixed exchange rates among the fifty states. What is highly questionable is the way in which monetary integration was pursued: with minimal concern for fiscal integration, so that individual European nations lost their currency sovereignty while no federal sovereign fiscal institution was created.

The European process of integration has always been characterised by the crucial role played by political factors. The creation of a currency area made up of significantly heterogeneous countries was an essentially political decision, and not the spontaneous outcome of their economic convergence. This followed from the idea that monetary integration itself would promote the degree of economic convergence among countries required for an efficient currency area. In this way, Europe could become one of Mundell's OCAs.

However, while recognising the importance of the political dimension, Europe overlooked the importance of the role of the state and politics in the process of integration. First, Europe eschewed the role that fiscal policy at the federal level can play in the process toward the creation *ex post* of an OCA. In a situation in which the countries that adopt a common currency are heterogeneous and characterised by low flexibility of prices and wages as well as low mobility of the factors of production, the risk of asymmetric shocks can be dealt with by the creation of a fiscal counterpart to the single central bank.

The process of monetary integration in Europe relied on a theoretical stance that sees fiscal policy as distorting and ineffective in the long run. More generally, state interventions are seen as something to avoid to the maximum possible extent in order not to disturb the spontaneous working of the economy. As a result, Europe adopted a quite unique process. The strong link between the state, which has fiscal authority, and the creation and administration of money was weakened to the point of having a central bank that is totally independent of national states and which has no fiscal counterpart.

In general, the concept of an independent central bank is both flawed and ambiguous, but in the concrete European experience the concept came true. The European Central Bank (ECB) is totally independent both with respect to the fixing of its objectives and with respect to the policies adopted to realise them. In this framework, individual national states are constrained because they cannot freely use fiscal instruments to affect output and employment. Europe seems to be entrapped in a vicious circle. The ECB's anti-inflationary stance and the European governments' requirement to be 'fiscally responsible' produce substantial stagnation. 'Cautious' (or restrictive) monetary and fiscal policies discourage the growth of investment and aggregate demand in general.

The endogenous nature of budget deficits means that slow growth impedes tax generation, causing deficits to widen. This makes it even more difficult to stay in line with the fiscal parameters, and hence, there is a further negative impact on demand.

In this context, the European Union relies on foreign demand as an engine of economic growth. But this gives rise to a circularity: each member state tries to increase net exports, both with other EMU nations and with the rest of the world, in part by trying to become a low-cost producer. As exchange rates are fixed with the rest of the EMU, the only alternative is to maintain or reduce wages and prices within the member state, or achieve higher productivity growth. The former adds more pressure for fiscal austerity and slow growth.

The euro experience provides a negative lesson for those advocating 'go it alone' approaches to pegged exchange rates, even in the case of large trading blocs with substantial international power.

Conclusion

The world has changed tremendously since the 1940s when the Bretton Woods system was formed. For one generation, that system performed reasonably well, with fixed but adjustable exchange rates. However, it was developed for a world in which capital flows were controlled, relatively small, and dominated by official flows (from the IMF, the World Bank and the US Marshall Plan). This after all, can explain why Keynes' own plan largely ignored the role and effects of capital movements on exchange rates. Further, even trade in goods and services was quite restrained, in part due to the overwhelming post-war dominance of the US.

Over time, the US lost ground as Europe recovered and Asia became a major producer. Likewise, private capital flows grew gradually, and then in a torrent, partly due to technological change and partly due to 'neo-liberal' policy that sought to free financial markets. Even before much progress was made in that direction, the Bretton Woods system collapsed. While some still nostalgically call for a return to a fixed exchange rate system based on Keynes' Bancor Plan, current economic and political trends make this highly improbable. Nor can most nations individually adopt fixed exchange rates, when speculative attacks can break virtually all pegs, except for a few modern mercantilist nations that have accumulated large hoards of dollars.

Pegged exchange rates remove an important degree of freedom, holding domestic fiscal and monetary policy hostage to the exchange rate. A floating rate allows for greater domestic policy independence, providing fiscal and monetary policy space. However, it must be stressed that adopting a floating exchange rate is no panacea. It is only a necessary condition for gaining policy independence. It does not by itself ensure either enlightened use of this policy independence or an easy path to growth and development. In the current world situation, floating exchange rates are a *necessary but not sufficient* condition for the implementation of policies that can promote more growth, employment and welfare. Ultimately, the adoption of such policies is contingent on the ability and willingness of social, political and economic actors to do so.

Note that a flexible or floating rate regime is not necessarily a 'free float' system. A flexible exchange rate can leave some room for discretionary intervention. Fiscal and monetary policy as well as official transactions in exchange markets can still be used to 'manage' exchange rates in some circumstances. In particular, interventions might be called for in the case of rapid revaluations to ease the competitive pressures arising from an overvalued currency. However, achieving domestic, internal, stability would be the primary goal of policy, with full employment the most important domestic policy objective.

Floating exchange rates give nations one more degree of freedom but of course they also imply some costs. Among such costs there can be a larger degree of uncertainty due to the possibility of some volatility of exchange rates and terms of trade, as well as the costs of the possible triggering of inflationary processes deriving from a large depreciation of the national currency which makes the price of imports increase. From this point of view, greater stability and independence could perhaps be achieved by some combination of floating exchange rates with capital controls and trade policy, especially in the case of developing nations. These factors would make it easier to adopt managed exchange rates. However, the question of how to introduce effective capital controls in the present world situation remains open.

Keynes' plan relied on the existence of some form of international governance. If it became possible to construct international institutions to promote economic growth by focusing attention on financing development, this could also improve international economic performance by making it possible to give exchange rates and terms of trade more stability.

In this perspective, the European experience is a telling example. In principle, monetary integration can be the correct response to the need for more stable external conditions in European countries. However, the present European arrangements, although they give stability to exchange rates, do not work effectively to guarantee more growth and higher levels of employment. The basic reason is that there is no supranational institution that plays the role played by the governments of sovereign nations. In other words, the EMU does not work satisfactorily for the same reasons that a world regime of fixed exchange rates cannot function well without a supranational institution with tasks analogous to those envisaged by Keynes and others. There is little political will to submit to such a supranational institution, and as such, a Bancor-type plan is not feasible.

31.6 Environmental Sustainability and Economic Growth

We started our study of macroeconomics in this textbook by noting that the main macroeconomic policy goals are full employment and price stability.

A central idea in both microeconomics and macroeconomics is efficiency; getting the best out of what you have available. At the macroeconomic level, the 'efficiency frontier' is normally summarised in terms of full employment, the situation in which all available labour resources are being productively deployed.

We have learned that the concept of full employment is hotly contested among different schools of thought in macroeconomics, but this does not negate the fact that the attainment of full employment – using our macroeconomic resources to the limit – remains a central focus of macroeconomic theory and policy. The debate is about what that limit actually is. However, this debate is framed by economists in terms of the extent to which there are structural impediments which might mean our conception of full employment is associated with higher rates of unemployment than we would like.

We learned in Chapters 20 to 24, that capitalist monetary economies are prone to deliver mass involuntary unemployment as a result of a lack of effective demand. The solution to involuntary unemployment involves increasing the level of effective demand so that it is consistent with the level that is required to employ all those willing and able to work and is filled with public net spending. This can be achieved by increased government spending to ensure that the shortfall in non-government spending relative to the full employment level of demand is filled. In addition, the government can also stimulate non-government spending in a number of ways, including tax cuts, interest rate cuts, and investment and export incentive schemes.

This suggests that sustaining the economically and socially desirable goal of full employment requires continuous growth in aggregate demand and real GDP as populations expand.

In Chapter 4, we learned that the conventional market-based measures of national income as indicators of well-being are flawed in several ways. First, many activities that nurture well-being (for example, domestic work, caring for our children) are not counted as economic activity, unless a payment for service is made. Further, any production that is sold in the marketplace will add to the conventional measures of GDP. So a society is considered to be 'growing' and doing better if it produces increasing quantities of military weapons, which then wreak havoc during times of conflict. Similarly, a major environmental disaster, such as the 2010 Deepwater Horizon oil spill in the Gulf of Mexico, boosts economic growth through its clean-up operation, even as it devastates the local marine environment.

We also learned that national income measures of well-being largely ignore distributional issues. How might we feel about two economies which are recording the same growth rates, but in one, the vast majority of the income growth is secured by a small minority and the rest of the population live in poverty, whereas in the other, the population broadly shares in the increased real income?

Conventional real GDP measures also ignore the costs of increasing depletion rates of non-renewable natural resources. The mining or forestry industries, for example, boost economic growth but the environmental damage that is left behind is not considered because the firms involved do not count the damage as a cost of production. The growth of industrial production will be 'good' for real GDP growth, but the associated pollution of the land, water and air that results from the activity might be damaging for our health, and ultimately, undermine the capacity of the economy to produce as the natural systems die. There is growing evidence that unsustainable farming practices are reducing the amount of available productive land and destroying waterways, but still count dollar for dollar in our measures of economic growth.

It is clear that the capitalist system is not only prone to creating mass unemployment, but it also grows on the back of environmental degradation which is destroying our natural capital. This would appear to require a shift in our 'growth at all costs' approach to economic policymaking.

Matthew Forstater wrote that:

> *Environmental degradation in the form of unsustainable rates of natural resource depletion and excessive pollution of land, air, and water is characteristic of modern capitalist economies. Humanity now faces significant challenges in the form of both local ecological crises and global environmental problems, such as ozone depletion, global climate change, biodiversity loss, soil erosion, and deforestation* ... (2003: 386)

The question that arises is whether the desire to maintain full employment and the requisite growth that is associated with that policy goal is consistent with environmental sustainability, given the increasing evidence that anthropogenic global warming and resource depletion is endangering the health of the natural environment upon which our economic and social settlements depend. While full employment appears to be a necessary social and economic goal, can we reconcile it with the obvious need to ensure that our natural environment is also sustained? Even if it were possible to expand aggregate demand enough to promote sufficient growth to keep pace with labour force growth and productivity growth, and to mop up the huge stocks of long-term unemployed, how could the natural ecosystems, which are already under great strain, cope?

While the full treatment of what constitutes environmental sustainability is beyond the scope of this textbook, some useful observations can be made. For further discussion of what constitutes environmental sustainability see Lawn (2001) and Forstater (2003).

The important point that emerges from that literature is that growth is not necessarily good or bad *per se*. We can certainly refine our measures of growth to reflect the shortcomings noted above. At a minimum, this requires us to include all the costs of production in our measures of economic activity. Our revised net measures of economic growth will then ensure that we understand whether the changing scale of economic activity is advancing our overall well-being or not. It is entirely possible that the conventional measure of real GDP might show a slowdown in growth (which we would currently interpret as a problem) while new improved measures of real GDP (net of costs) would show an improvement in sustainable growth.

But we can also engender growth that will be sufficient to ensure that all those who want work can find a job at decent pay and conditions and still satisfy the requirements of environmental sustainability.

Clearly, this suggests that it will be necessary to change the composition of final output toward environmentally sustainable activities. It is not increased aggregate demand *per se* that will be necessary to sustain full employment, but increased aggregate demand in certain areas of activity.

Policymakers who are intent on eliminating the waste of human potential must also pursue the broader aim of preserving our natural capital and minimising the waste that arises from resource extraction. That is, our notion of a 'macroeconomic efficiency frontier' thus has an extra constraint on it: the need to protect our natural capital.

Phillip Lawn defines this in terms of an "optimal macroeconomic scale",

> *where the physical scale of a nation's macroeconomy and the qualitative nature of the goods with which it is comprised maximises the sustainable economic welfare enjoyed by its citizens. The notion of optimal macroeconomic scale is a crucial one because it enables one to understand how the nation can achieve [sustainable development] ... without the perceived need for continued growth.* (2001: 1–2)

Several questions arise when considering the concept of an 'optimal macroeconomic scale'. First, what are the benefits of creating and maintaining wealth derived from economic activity? Second, what does this cost in terms of depleting the available environmental services?

Natural capital is identified by Lawn (2001: 4) as "the original source of all economic activity" because it "is the sole source of low-entropy matter energy and the ultimate repository of all high-entropy wastes". Economic activity generates real income (which is the satisfaction derived from consuming goods and services produced) but also imposes social costs on those who produce these goods and services.

Lawn notes that we have to endure personal costs such as the disutility of work and the stress of commuting as part of the process of generating these economic benefits and the calculation of what he calls 'net psychic income' reflects the human costs and benefits of economic activity.

While calculating the human costs of economic activity, we also have to take into account the costs of depleting the available environmental services. Extracting these services (the low-entropy matter energy) creates waste (high-entropy matter energy) which has no further use. This waste depletes our natural capital and is the ultimate resource cost of economic activity.

Trying to define the costs of natural resource depletion is fraught because the biosystem is a living entity and economists are unable to define the use point beyond which it dies.

Sustainable net benefits are the difference between the net psychic income and the environmental costs of economic activity. The maximum macroeconomic scale for a nation occurs when the sustainable net benefits are zero. Beyond this physical scale of production, the environmental costs exceed the net psychic income derived from economic activity.

The point of maximum sustainable net benefits is the optimal macroeconomic scale, because the difference between the net psychic benefits and the environmental costs are at their greatest. The macroeconomic efficiency frontier thus is defined by the juxtaposition between the net human benefits and the environmental costs required to create those net human benefits. The full employment goal then needs to be expressed not just in terms of the total number of jobs required, but also the type of jobs and the activities these jobs will be engaged in.

Growth is necessary for employment as the population grows, but it cannot be unfettered and left to the market. It must be a carefully guided growth with new tools that are not normally used by economists to help governments decide on what their contribution should be and how they should regulate the contribution of the non-government sector. Consideration of that will lead to a rather dramatic reshaping of our conception of productive work, which is currently narrowly defined in terms of 'gainful' paid endeavour in pursuit of (private) profit.

References

ABS (Australian Bureau of Statistics) (Various years) Commonwealth Fiscal Paper 1. Available at: http://www.abs.gov.au/, accessed 14 June 2017.

Bernanke, B.S. (2004) "The Great Moderation", Board of Governors of the US Federal Reserve System, at the meetings of the Eastern Economic Association, Washington DC, 20 February.

BLS (Bureau of Labor Statistics) (Various years). Available at: https://www.bls.gov/, accessed 12 October 2018.

Board of Trustees for the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds (2016) *Annual Report*, US Government Printing Office, Washington. Available at: https://www.ssa.gov/OACT/TR/2016/tr2016.pdf, accessed 13 June 2017.

Boughton, J.M. (2001) "Was Suez in 1956 the First Financial Crisis of the Twenty-First Century?" *Finance and Development*, 38(3), September. Available at: http://www.imf.org/external/pubs/ft/fandd/2001/09/boughton.htm, accessed 14 June 2017.

Dantas, F. and Wray, L.R. (2017) "Full Employment: Are We There Yet?", Levy Public Policy Brief, No. 142, February.

Executive Office of the President of the United States (2016) "The Long-Term Decline in Prime-Age Male Labor Force Participation", June. Available at: https://obamawhitehouse.archives.gov/sites/default/files/page/files/20160620_cea_primeage_male_lfp.pdf, accessed 13 June 2017.

Forstater, M. (2003) "Public Employment and Environmental Sustainability", *Journal of Post Keynesian Economics*, Spring, 25(3), 385–406.

Keynes, J.M. (1980) *The Collected Writings of John Maynard Keynes: Activities 1940–1944. Shaping the Post-War World: The Clearing Union*, London: Macmillan.

Lawn, P.A. (2001) *Toward Sustainable Development: An Ecological Economics Approach*, Boca Raton, FL: CRC Press.

Mundell, R.A. (1961) "A Theory of Optimum Currency Areas", *American Economic Review*, 51(4), 657–65.

Pigeon, M.A. and Wray, L.R. (2002) "Demand Constraint and the New Economy" in P. Davidson (ed.), *A Post Keynesian Perspective on Twenty-First Century Economic Problems*, Cheltenham: Edward Elgar, pp. 158–94.

Sardoni, C. and Wray, L.R. (2007) "Fixed and Flexible Exchange Rates and Currency Sovereignty", Levy Economics Working Paper No. 489, January.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER 31 APPENDIX 1: CASE STUDY 1 – ECONOMIC GROWTH: DEMAND OR SUPPLY CONSTRAINED? THE US, 1975 TO 2007

Introduction

In this case study, we will examine economic growth in the United States over much of the past half century.[1] The period from the early 1970s through to the mid-1990s was characterised by substandard growth and relatively high unemployment that stood in stark contrast to the earlier post-war period. By the early 1990s, many economists had concluded that growth was constrained by supply side factors, namely fewer technological advances, lower attachment to the labour force by young workers, and the poor skills of job seekers.

Surprisingly, all this changed in the second half of the 1990s as the 'New Economy' surged ahead with dotcom innovations while unemployment fell close to the low rates that had last been experienced in the 1960s. While the economic growth slowed with the dotcom crash and a recession occurred at the end of the 1990s, growth resumed in the early 2000s with the synergies of the commodities, real estate, and stock market bubbles. The earlier musings about 'speed limits' to growth, which originated on the supply side, were forgotten for more than a decade.

Indeed, between 1995 and 2007 faster growth was accompanied by rising productivity which temporarily attenuated the talk about supply side constraints. Many observers were claiming that we had entered a 'new era' of higher growth, lower unemployment, stable prices, and financial stability, that is the 'New Economy' and the 'Great Moderation' as described by two Chairmen of the US Federal Reserve, Alan Greenspan and Ben Bernanke (2004), respectively.

In the orthodox approach, long-run growth is supposed to result from a combination of growth of the factors of production (quantities of labour and capital) and improvements to productivity (a function of education and training of workers and technological advances). Note that these are all supply side variables. It is presumed that demand factors do not matter in the long run because the ups and downs of the business cycle will net out on average.

We will see that it is a mistake to ignore demand factors for two reasons. First, if an economy suffers from chronically insufficient demand, then average long-run growth will be lower than it would have been with higher aggregate demand. Second, low demand means that both labour and capital are underutilised. Low labour demand promotes hysteresis, with new entrants to the labour force being unable to find jobs and being denied relevant skills (and the socialisation associated with stable work patterns) and redundant workers facing skill obsolescence and developing habits and attitudes not conducive to hiring (see Chapter 18). Furthermore, firms have little incentive to invest in training if unemployed workers are plentiful. Firms may simply raise their hiring standards to take advantage of slack labour markets. The inducement to invest is weak when existing capacity is underutilised. This leads to a slower rate of technological advance.

Thus, mainstream economists have been reluctant to recognise that the real problem has been insufficient aggregate demand in periods of weak growth, not supply side limitations. This mistake in diagnosing the problem leads to the wrong policy prescription. The mainstream economists typically advocate the adoption of restrictive policy, believing that the economy has come up against supply side constraints long before full employment. This compounds the problem by keeping aggregate demand too low to induce investment, technological advance, and skills upgrading of the workforce.

We now turn to the first slow growth episode (mid-1970s to mid-1990s), and to the so-called New Economy boom. In Case Study 2, we look at the post-GFC period, when another slow growth period renewed fears of secular stagnation.

In this case study, we choose to focus on GDP per capita and the employment rate rather than the levels of these variables (GDP and employment) because many people measure progress by the income growth that the average person enjoys over time.

Per capita GDP (constant price GDP divided by the total population (*POP*)) can be written as the product of the employment rate (workers (*N*) divided by total population) and productivity per worker according to the following formula:

(31.2) $GDP/POP = (N/POP)*(GDP/N)$

Note that here the employment rate is the ratio of employment to the total population, and not the conventional employment-to-working-age population ratio which is published by most government statistical agencies.

Expressed in growth terms, Equation (31.2) would tell us that, as a rule of thumb, the percentage growth in GDP per capita is the sum of the growth in the employment rate (*N/POP*) plus productivity growth.

Did the US economy suffer from secular stagnation from 1970 to 1995?

Discussion of the possibility of 'speed limits' to growth coming from the supply side was common in the US in the 1990s (until the 'New Economy' of the dotcom bubble boosted growth). Table 31.1 compares the per capita inflation-adjusted GDP of the major developed nations, with each country's per capita GDP indexed to 100 in 1970. Of note is the relatively rapid growth of Japan over the earlier part of the period, and of the UK from the late 1980s. Otherwise, per capita growth was similar among these countries, with the US in the middle of the pack of slower-growing countries. All nations experienced growth in productivity over the same period, with Japan easily the best performing nation. The US, Canada and Australia were at the bottom of the productivity growth list.

Figure 31.3 shows the change in employment rates and labour productivity of the major developed nations between 1970 and 1995. Remember that we are dealing with movements in the employment **rate** and GDP **per capita**. To help you better understand what is happening in Figure 31.3, it is significant that between 1970 and 1995, real GDP in the US grew by 115 per cent (on average 3.1 per cent per annum), whereas GDP per capita grew by 63.8 per cent (on average 2.1 per cent per annum). The discrepancy arises because US population growth was very robust during this period (on average one per cent per annum), which meant that there were more people to share in the growth (on average). This contrasts with what was happening in the European economies and Japan, where population growth was much lower.

Further, US employment was growing very strongly between 1970 and 1995 (58 per cent overall or an average of 1.9 per cent per annum) but the strong population growth meant that the employment rate only grew by 23 per cent (an average of 0.7 points per annum). The employment growth in the European nations (Germany excepted) and Japan was well below that of the Anglo nations, with Canada and the US leading the way.

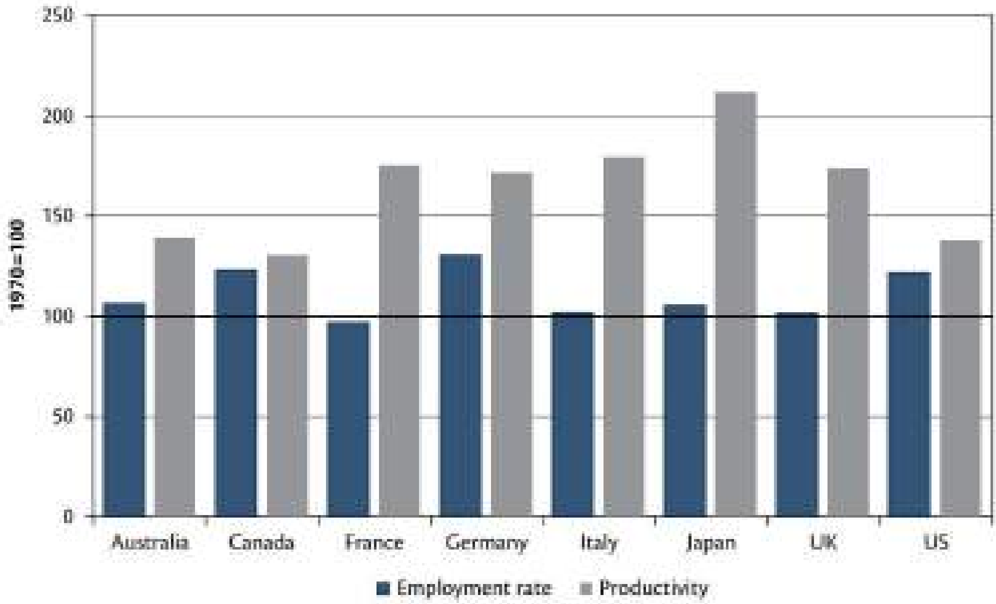
In the case of the US, the focus of this case study, the question arises as to whether the slower relative productivity growth was due to supply side factors. As discussed above, the mainstream argues that the tendency toward secular stagnation was due to slower improvements of technology, as well as to lower motivation and skills of

Table 31.1 GDP per capita 1960 to 1995, 1970=100, 2011 prices

	Australia	Canada	France	Germany	Italy	Japan	UK	US
1960	73.9	72.6	64.0	70.9	61.5	42.2	77.6	75.2
1970	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1980	113.3	129.4	135.2	131.0	138.5	137.8	121.3	123.1
1990	134.9	151.2	162.1	160.5	174.8	205.1	161.9	154.2
1995	146.0	155.6	168.6	181.8	186.9	216.6	176.0	163.8

Source: Data from Bureau of Labor Statistics – Table and calculations are authors' own.

Figure 31.3 Employment rate and labour productivity, 1970 to 1995, 1970 = 100



Source: Data from Bureau of Labor Statistics and authors' calculations.

young workers. It seems quite implausible that the US alone suffered from such maladies while Europe and Japan somehow escaped them. The implausibility of this argument is driven home by the fact that in the mid-1990s the US quite suddenly entered a 'new era' with a burst of what was said to be supply side-driven productivity growth (see below).

Is there another possibility? Perhaps the problem was demand side constraints rather than supply side constraints.

Using the rule of thumb that the percentage growth in GDP per capita is the sum of the growth in the employment rate (*N/POP*) plus productivity growth, Figure 31.4 shows how the long run increase in GDP per capita for the US was split between the changes in the employment rate and productivity in the years 1970 to 1995. On average, annual growth in GDP per capita was 2.1 per cent, while the employment rate rose by 0.8 points per annum, and productivity grew at 1.2 per cent per annum.

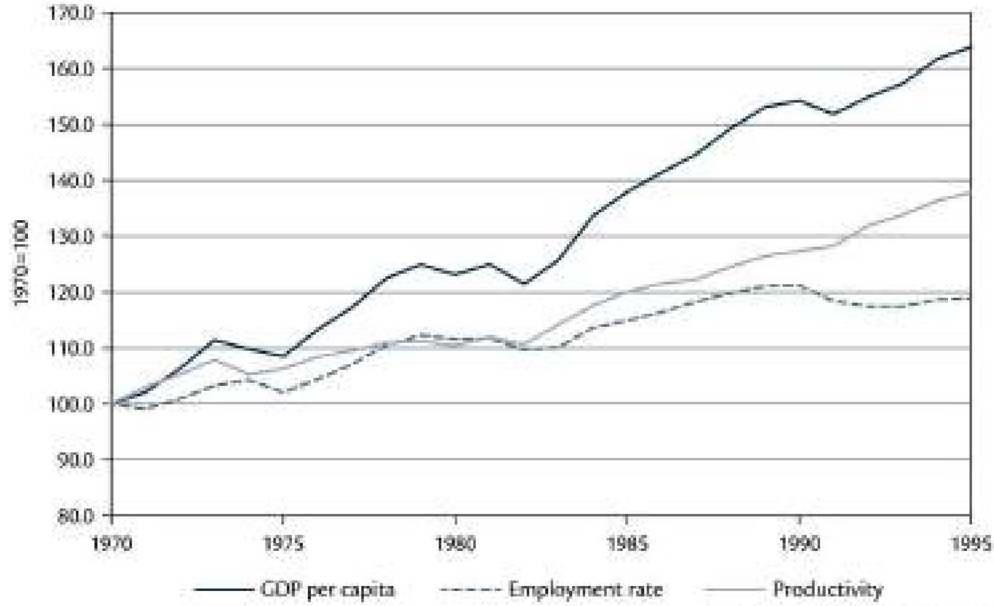
Equation (31.2) is an identity, which means that it is true as a matter of accounting. To give meaning to the movements in the components over time, we seek to understand the causal links between the components. Further, while economists and policymakers habitually impute great importance to measured productivity growth even over very short periods of time (that is, quarter over quarter), in practice, measured productivity growth by itself cannot tell us that supply side factors are the cause.

Why? The reason is that while supply side factors (for example, new technologies) can be important, none of the right-hand-side components of Equation (31.2) are exogenous. Employment growth (and hence changes in the employment rate) and productivity growth are driven by growth in GDP, which in turn translates into changes in GDP per capita for given population growth. There is thus a strong interdependency between the various terms in Equation (31.2).

Hence, we cannot easily say that faster productivity growth was driving growth in GDP (and GDP per capita) and that this growth was driven by supply side factors. Even if there were supply side factors in play, the higher GDP growth also would have elicited growth in the employment rate (given the population growth) and productivity growth.

What we can say is that the low relative productivity growth in the US was likely to result from weak demand. Further evidence to support that conjecture can be deduced from the fact that US unemployment rates were significantly higher over the period than they had been before the mid-1970s. Further, wages growth lagged

Figure 31.4 Decomposing the increase in US GDP per capita, 1970 to 1995, 1970 = 100



Source: Bureau of Labor Statistics and authors' calculations.

behind both inflation and productivity growth, implying a shift of the distribution of income toward capital and away from labour. This is another indication that labour markets were not tight, as loose labour markets kept wage growth low. Those low wages in turn, might have reduced the incentive for firms to undertake productivity enhancing investment in capital equipment and labour training.

All of this points to the importance of demand constraints rather than to supply constraints for the lower relative productivity growth in the USA during the period 1970 to 1995.

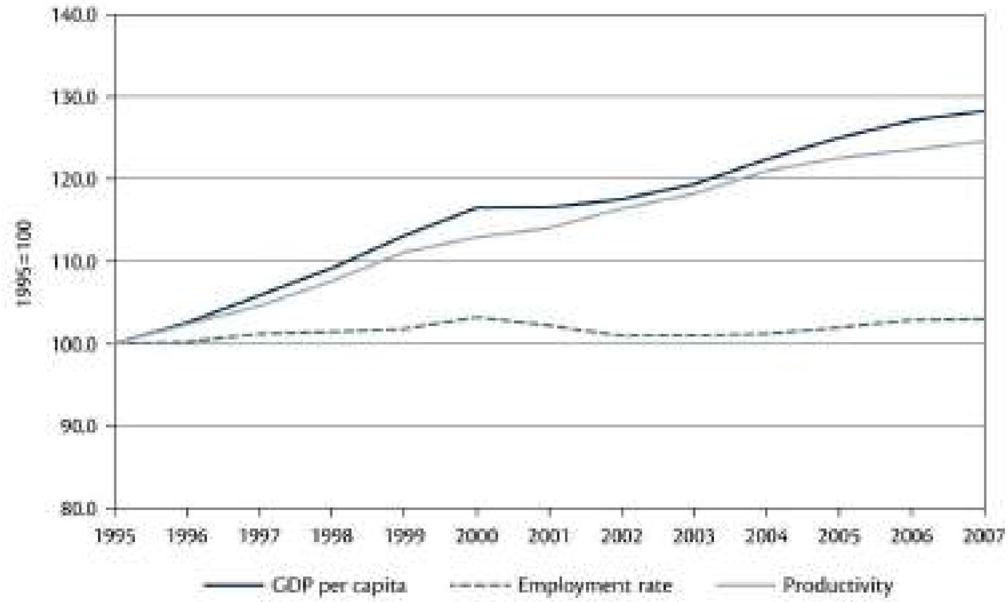
The 'New Economy' and the productivity miracle, 1995 to 2007

During the mid- to late 1990s boom associated with the Presidency of Bill Clinton there was much talk about new information technology that had allegedly made it possible to grow at higher rates. From 1995 to 2000, average real GDP growth rose to 4.30 per cent, while per capita real GDP growth managed to achieve a robust 3.11 per cent rate. This was called 'Goldilocks' growth: high enough to cause many economists to believe a 'New Economy' had emerged, and banishing talk of secular stagnation. However, when the New Economy faltered in the 'dotcom' crash at the end of the decade, a recession lowered growth of real GDP to just under one per cent for 2001, followed by a slow 'jobless' recovery. It was only the run-up of the housing and commodity bubbles that brought real GDP growth for 2002–6 to 3.15 per cent, and real per capita GDP growth to 1.75 per cent for that recovery period. The economy began to slow and then crashed into what would become the Global Financial Crisis (GFC) beginning in 2007.

As it turned out, the average annual rates of growth in GDP per capita over the whole period between 1995 and 2007 were identical to that of the earlier period (2.1 per cent per annum). The difference between the periods was that productivity growth increased from 1.2 per cent per annum on average to around 1.8 per cent, while the growth in employment fell from 1.9 per cent per annum (1970 to 1995) to 1.3 per cent per annum average, just above the annual increase in the population. As a result, the employment rate only rose by 0.3 per cent per annum in this period (down from an average of 0.8 per cent in the earlier period).

As Figure 31.5 shows, the rise in GDP per capita was accompanied by relatively stronger increases to productivity even with a mostly static employment rate. The data would appear to support the claim that the growth in GDP per capita during the dozen years between 1995 and 2007 can be attributed to higher aggregate demand, which reinforced the productivity growth arising from the supply side. This is despite weak employment growth.

Figure 31.5 Decomposing the increase in US GDP per capita, 1995 to 2007, 1995 = 100



Source: Bureau of Labor Statistics and authors' calculation.

The two features of the period were:

- The private sector was willing to spend far more than its income and ran up large deficits and accrued growing levels of debt until the nation collapsed in the GFC. The credit boom ensured that growth in aggregate demand would be strong.
- At the same time, the technology boom and the accompanying higher productivity growth meant that the growth in employment was attenuated, and with continuing population growth this meant that the era of a rising employment rate had come to an end.

The 'productivity boom' did not last for long however, as the economy slowed with the bursting of the dotcom bubble and then crashed into the GFC. In the next case study, we turn to the sluggish recovery from the deep recession, which led to renewed claims of secular stagnation.

CHAPTER 31 APPENDIX 2: CASE STUDY 2 – THE RETURN OF SECULAR STAGNATION? US LABOUR MARKETS AFTER THE GLOBAL FINANCIAL CRISIS[4]

During the deep recession induced by the GFC of 2007, employment and economic growth in the US plummeted, followed by a very slow recovery over the next decade. In these circumstances, it was clear that the high unemployment and low growth were due to insufficient aggregate demand.

However, rather than promoting and then sustaining full employment, the US Federal Reserve Bank made it clear that it was worried about exhaustion of the supply of employable labour, so it would continue to raise interest rates to fight the inflation expected from low unemployment rates.

Further, well before unemployment had fallen to pre-recession levels, some economists were already beginning to argue that slow growth and relatively low employment were 'the new normal'. They argued that we would have to lower our expectations and accept what is called 'secular stagnation', which means sustained sluggish growth.

While there is consensus that the economy was constrained by insufficient demand immediately after the recession, many economists continued to argue that longer-term growth is supply constrained. The main limiting factors are said to be the lower quality and insufficient quantity of employable labour in conjunction with a lower rate of technological change. The falling growth in the labour force was attributed to low birth rates, and an ageing population. In addition, while the US benefitted from rising labour force participation by women from the 1960s through the 1990s, that source of additional labour had been exhausted by the 2000s. Furthermore, structural changes in labour markets reduced the demand for workers with lower skills and education. Too many young people do not complete enough schooling to meet the needs of modern production processes. Finally, the technological wave created by the computer age had run its course, and no new breakthroughs were on the horizon that would unleash major productivity advances.

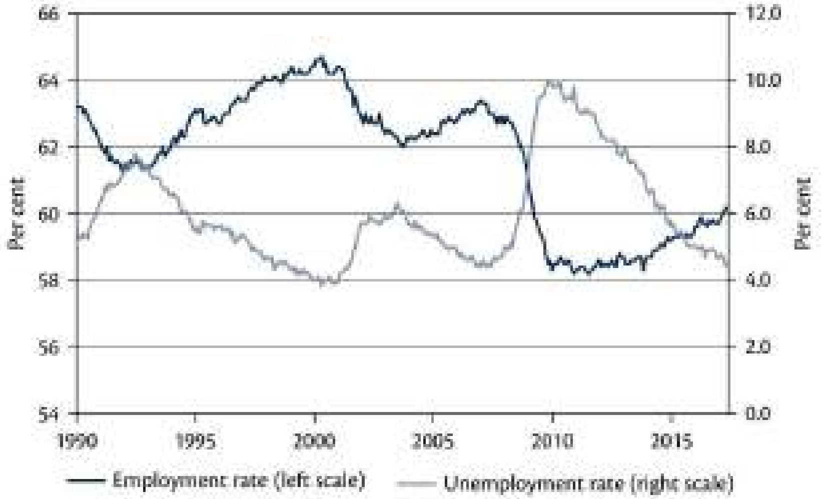
Unlike the secular stagnation conjectures in the 1970 to 1995 period, which had focused on the slowdown in productivity growth, the focus in this later period centred on the supposed constraints on employment growth.

Figure 31.6 shows the employment-to-working age population ratio (total employment divided by the working age population 'employment rate') and the official unemployment rate. We are using the employment-to-working age population ratio in this section, instead of the employment rate that we focused on earlier, because it allows for a quarterly comparison with the movements in the unemployment rate.

The ratio of employment to working age population remained well below the levels that existed before the crisis. In fact, since 1990 the pace of recovery of the employment-to-working age population ratio has been painfully slow after each recession. Perhaps most concerning is an increasingly asymmetric response of the unemployment rate (unemployment as a percentage of the labour force) and the employment-to-working age population ratio over the course of the cycle.

This asymmetry has two features which are illustrated in Figure 31.6. First, in a downturn the unemployment rate rises sharply while the ratio of employment to working age population falls sharply. In the recovery, both are slow to revert to previous levels. Second, while both rates change sharply in the downturn in opposite directions, the unemployment rate declines more quickly than the ratio of employment to working age population improves. The empirical time series relationship between the employment-to-working age population ratio and the unemployment rate is mediated by changes in population, the dependency ratio and the rate of labour force participation. Further discussion of this relationship is beyond the scope of this textbook.

Figure 31.6 US employment-to-working age population ratio and unemployment rate, January 1990 to April 2017



Source: Bureau of Labor Statistics

While the official unemployment rate has fallen back to pre-GFC levels, broader measures of labour underutilisation reveal that the recent recovery is also characterised by substantial numbers of workers who can only find part-time work when they want full-time jobs. The presence of significant **underemployment** means that some additional aggregate demand can be met by incumbent workers working longer hours, without additional employees being required.

Figure 31.7 plots the official unemployment rate against the 'unofficial' U6 measure,[5] which includes not only those who are marginally attached to the labour force, but also the underemployed.

As of April 2017, the U6 unemployment rate was 8.6 per cent. There were still 7.1 million people unemployed, 5.3 million people employed part-time for economic reasons, and 1.5 million people marginally attached to the labour force. The vast majority of the marginally attached were discouraged (could not find a job), or out of the labour force for other factors that made participation difficult, including childcare and transportation difficulties. This is not surprising given the difficulty of accessing affordable, available and adequate childcare and public transportation in the US.

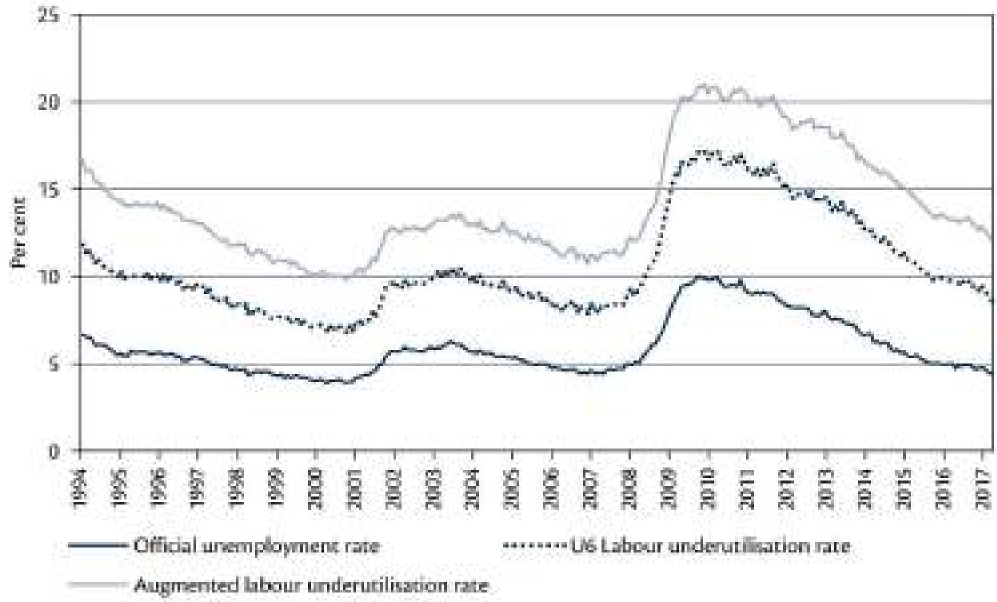
However, even the U6 measure of unemployment is likely to understate the challenges, because the BLS statisticians consider 'marginally attached' to include only those who have searched for employment in the previous year. In April 2017, according to the BLS, there were 5.7 million people outside of the labour force who reported wanting a job now, a high proportion of whom had not searched for work in the previous year. Taking these people into account, a more comprehensive measure boosts the rate of idle labour to 12 per cent, a measure labelled 'Augmented labour underutilisation rate' in Figure 31.7. This means that there were approximately 20 million people who would like a full-time job yet were unable to find one in the US in April 2017.

The typical explanation for declining US labour force participation is age demographics, combined with changing characteristics of the American labour force and other structural forces. Remember that the aggregate labour participation rate is a weighted average of the participation rates of the individual age cohorts, with the weights being the proportion of each cohort in the total labour force. Further, older workers have much lower participation rates than prime age (25 to 54 years) workers.

Thus, even though the labour force participation for those aged 55 and over started rising in the late 1990s, the rising proportion of the older workers with lower participation rates and the declining weight of prime age workers, meant that the overall participation rate started to decline from early 2000, and that trend has continued.

One of the puzzles associated with the trend has been the concomitant decline in prime age participation rates. The US now has the lowest labour force participation rates for prime age workers among OECD countries. The declining participation of this cohort has been one of the important long-run drivers in the decline of overall

Figure 31.7 Measures of labour underutilisation, USA, 1994 to 2017, per cent



Source: Bureau of Labor Statistics and authors' calculation.

Note: The U6 labour underutilisation rate is calculated as the ratio of unemployed workers plus employed part-time for economic reasons plus marginally attached to the labour force over the civilian labour force plus the number of marginally attached workers. The augmented labour underutilisation rate is the sum of the U6 rate and BLS series LNS15026639 Not in the Labour Force, Want a Job Now.

labour force participation, and one which brings into question the whole 'ageing workforce' explanation for the declining US participation.

It is hard to make the argument that excessive labour market regulation, or generous social safety nets can explain declining participation in the US as compared with other OECD countries, since those countries generally have labour markets that are more regulated, and social safety nets that are more extensive and comprehensive.

Some have attributed the decline of labour force participation to 'social shifts' such as changes in personal preferences as more people aspire to a better work/life balance (spending more time with family, on leisure interests, and so on), or to seek higher educational achievements. While the decline of labour force participation for those between the ages of 16 and 24 reflects the fact that Americans are spending more years in school, this cannot explain why Americans (especially men) of prime working age have withdrawn from the labour force.

While the social shifts explanation sounds plausible, it is unlikely that a large number of Americans are voluntarily leaving the labour force in accordance with personal preferences. In fact, the number of those not in the labour force who report **not** wanting a job now has declined for age groups 16–24 and 25–54. Further, the historical trend for married couples with children under 18 shows a significant increase in the percentage of families in which both parents are employed, from 25 per cent in 1960s to almost 61 per cent in May 2016. Neither of these trends seems consistent with the 'lifestyle changes' argument.

The number of males of prime working age who are neither employed nor looking for a job has more than doubled over the past 50 years. However, 'personal choice' to spend more time with family while a spouse works seems to explain little of this decline because fewer than 25 per cent of prime age people who are not participating in the labour force have a working spouse, and nearly 36 per cent of them were living in poverty (Executive Office of the President of the United States, 2016). Furthermore, the decline in participation among prime age males without children is more than twice that of prime age males with children.

While the ageing effect on participation is not denied, the deficient aggregate demand in the aftermath of the GFC has created a discouraged worker effect. This is the tendency for the long-term unemployed to be classified as being outside the labour force because, in the face of scarce employment opportunities, they cease to actively seek work. In other words, a proportion of the decline in the US participation rate since 2000 is because hidden unemployment has risen.

Conclusion

Despite the widespread belief that supply side factors impinge on the ability of the economy to grow faster over the long run, the evidence is that the US has operated with substantial labour market slack since the mid-1970s. Lack of sufficient aggregate demand has also caused slower productivity growth over most of this period. If anything, the slack has risen in the aftermath of the GFC, in part due to the discouraged worker effect.

This slack, and accompanying downward pressure on wages, reduces the incentive for firms to invest in labour-saving production techniques. That in turn, reduces the incentive to innovate in order to raise labour productivity.

In other words, lack of effective demand produces a vicious cycle of perverse disincentives that together keep aggregate demand and hence economic growth lower than they might have been. Unfortunately, policymakers misread this as the existence of supply side constraints on growth because they think that measured productivity is largely a supply side phenomenon.

The substantial pool of idle labour in the US could be put to work to increase aggregate demand and economic growth, which would likely be beneficial for productivity growth. While part of the policy solution is to encourage private demand, there is also room for more government spending. Since the private sector will hire the most employable workers first (those with more education, training, and work experience), it is necessary for the government to take up the slack that remains even in a robust expansion. The best way to do this is with the Job Guarantee.

CHAPTER 31 APPENDIX 3: THE US SOCIAL SECURITY AND MEDICARE SYSTEMS

The US social security and Medicare systems are organised into two separate trust funds, each having six trustees. Two public trustees (one Democrat and one Republican) scrutinise the advice of the other four, namely Secretary of the Treasury (the Managing Trustee), the Secretary of Health and Human Services, the Secretary of Labor, and the Social Security Commissioner.

The Trustee's Reports are annual legal requirements under the US Social Security Act. The reports document the financial status of each trust fund to the US Congress. Detailed projections of the revenues and outlays for each of the funds are generated, based on actuarial and economic assumptions and modelling.

The Old Age and Survivors Insurance Trust Fund is the accounting device used by the US Government to pay future retirement benefits. Payments of disability support pensions are accounted for via the Disability Insurance Trust Fund. Collectively, these funds are known as the Old Age, Survivors, and Disability Insurance (OASDI) programme.

The US Social Security Trustees Report 2016 notes that:

> *Under the Trustees' intermediate assumptions, projected OASDI cost will exceed total income by increasing amounts starting in 2020, and the dollar level of the combined trust fund reserves declines until reserves become depleted in 2034. ... Projected OASDI cost generally increases more rapidly than projected noninterest income through 2038 primarily because the retirement of the baby boom generation will increase the number of beneficiaries much faster than the number of covered workers increases, as subsequent lower-birth-rate generations replace the baby-boom generation at working ages. ... [C]onsider that for the combined OASI and DI Trust Funds to remain fully solvent throughout the 75-year projection period: (1) revenues would have to increase by an amount equivalent to an immediate and permanent payroll tax rate increase of 2.58 percentage points...; (2) scheduled benefits would have to be reduced by an amount equivalent to an immediate and permanent reduction of about 16 per cent...; or (3) some combination of these approaches would have to be adopted. (Board of Trustees, Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds 2016: 4–5).*

This type of reporting is very common. It is based on the premise that the social security programmes will be unable to pay the benefits unless there is a positive balance in the trust funds. Further, it is argued that the so-called funding gap will grow over time as the number of fund dependants grows. The US Social Security Trustees Report 2016 said that in 2038, social security benefits (OASDI) will be about 16.61 per cent of taxable payroll, with the Trust Fund having been exhausted by 2034. Clearly, these trust funds are the final responsibility of the US government, which can always ensure benefits are paid (in nominal terms). That capacity is intrinsic to the fiat monetary system. To obscure that intrinsic capacity, successive generations of politicians have placed a series of restrictive administrative layers, which they use to justify their claims that the government is financially constrained. These claims reflect the ideological distaste for government programmes of many conservative politicians. The organisation of the trust funds and the way they are funded exemplifies the way the public are misled into thinking their future pension and health entitlements are in doubt as a result of the alleged inability of the US government to honour all its liabilities. For example, the accounting smokescreen manifests through the US government's decision to allocate payroll tax and income tax revenue to the revenue side of these funds. Such an explicit gesture, which is totally unnecessary from a financial perspective, gives the impression to the general public that: (a) their taxes are being put to some specific use; and (b) that without those taxes funding those uses, there would be no capacity to provide the services.

Of course, both perceptions are false. Taxpayers do not fund anything in a fiat monetary system and the US government does not need any funding to provide whatever goods and services to the non-government sector it chooses.

In the US social security system, the deception is very powerful. Should the current revenues be insufficient to match current obligations, then the prior trust fund balances (which mainly represent past surpluses) are used. These past surpluses are invested in Treasury bonds.

It looks to be an elaborate mechanism for the government lending to itself and pretending to raise revenue that is necessary to 'fund' pension and Medicare entitlements. In fact, the US Government (via Congress) can always pass legislation which transfers 'general funds' into 'trust funds'.

While in an accounting sense, these trust funds can become insolvent because of the way they are currently constructed and accounted for, there is never a possibility that the US government would be unable to provide the requisite US dollars to pensions and healthcare, should it wish to do so. The summary facts are:

- The US Government can pay all benefits in each year with the stroke of a computer key. The only question would be whether the nominal transfers were possible in real terms, that is, will there be enough real goods and services for older Americans to enjoy in 2085? As argued above, high levels of private and public capital formation will enable these demands for goods and services to be met.
- There is no need to eliminate any deficits in these contrived trust funds, now or later. They are accounting fictions.
- There is no impending financial crisis in the US Social Security Trust Fund. The US government can always fund any entitlements at any time.
- The billions that go to social security each year will not make it harder to either find money for other government programmes or require large and ongoing tax increases. The US government is not financially compromised in meeting spending on Y by spending on X. It might be politically compromised, but that is nothing to do with the underlying features of the monetary system and the capacity that the US Government enjoys as the monopoly issuer of the currency.

Younger workers do not have to worry at all about the fact that their parents and grandparents are enjoying benefits from the funds. The largest problem facing younger workers is the entrenched unemployment and the rising duration of unemployment brought about by the failure of the US government to expand its deficit enough to support appropriate levels of job creation. This unemployment will cause intergenerational disadvantage and trying to cut spending now to fix up social security will worsen the situation for younger workers.

Endnotes

1. The Marshall Plan (so named for the US Secretary of State, George Marshall, who first advocated it) was an American initiative that from 1948 to 1952 provided over \$13 billion (c. \$110 billion in 2016 US\$) in economic aid to Western European states to help rebuild their economies in the aftermath of the Second World War.
2. The US accomplished the same feat during the Second World, when short-term Treasuries paid 3/8 of one per cent even as the deficit-to-GDP ratio reached 25 per cent of GDP! Rates on longer-term government debt are determined more complexly, but it is important to note that it is the government's decision to issue longer-term government debt. It can always accomplish the objective of draining excess reserves by issuing short-term bonds, so it does not need to issue long-term bonds. Indeed, if it pays interest on reserve balances, it does not need to sell bonds at all.
3. This analysis draws on Pigeon and Wray (2002).
4. This subsection draws heavily on Dantas and Wray (2017).
5. The U6 rate is provided by the Bureau of Labor Statistics (BLS).

CHAPTER

32 MACROECONOMICS IN THE LIGHT OF THE GLOBAL FINANCIAL CRISIS

Chapter Outline

Learning Objectives

- Understand the key deficiencies of mainstream macroeconomics which explains why their supporters did not foresee the Global Financial Crisis.
- Acknowledge that MMT advocates did anticipate the GFC and recognise the design faults of the European Monetary Union.

32.1 Introduction

In Chapter 26, we looked at alternative theories of the business cycle, from Marx to Minsky. As we saw, there are two main differences between orthodox theories of the cycle and heterodox theories.

First, orthodox approaches generally blame exogenous factors such as an external 'shock' or a policy 'surprise' for the cycle that moves the economy away from its 'natural' equilibrium. Endogenous market forces then move the system back to equilibrium. Real Business Cycle theory is an exception because the economy is seen to be always in equilibrium, with technology 'shocks' being the driver that shifts the economy to a new equilibrium growth path. Second, the cycle is not fundamentally a *monetary* or *financial* phenomenon.

While it is true that some orthodox approaches (such as Friedman's Monetarism and the New Classical theory of Lucas) blame monetary 'surprises' for the cycle, the normal role for money in orthodoxy is simply to determine nominal values, so it is 'neutral'. Only if economic agents make mistakes does money become temporarily non-neutral. It is that short-run, non-neutrality that can cause a business cycle.

By contrast, in most heterodox theories, the cycle is *endogenous*, whereby money and finance play a key role. As Minsky would put it, the business cycle is a financial cycle. Following Keynes, many heterodox economists argue that the cyclical nature of capitalism arises because the economy is organised around making money. Money can never be neutral because 'more money' is the object of capitalism. When capitalists either experience problems making money profits, or even if they just expect that there could be problems, economic growth can falter. That compounds the difficulty, particularly because there are monetary debts that must be serviced out of monetary income.

Minsky argued that capitalism is a financial system, in which there is a complex network of financial commitments. If there is a break in one of those commitments, for example one debtor defaults, this can bring down another debtor who relies on receiving payments from the economic entity that failed. While a cyclical downturn is not always accompanied by a financial crisis (and vice versa), those downturns that are accompanied by financial crises tend to be worse.

In this chapter, we will continue our investigation of the Global Financial Crisis (GFC) that began in 2007. The GFC was the worst global economic crisis since the Great Depression of the 1930s; in both cases, the economic downturn was accompanied by a severe financial crisis. Note that while both of these crises affected much of the world, in both cases there were large portions of the globe that escaped. Most notably, the USSR was not greatly affected by the Depression of the 1930s, and the GFC did not affect China's growth very much during the late 2000s.

Before the Great Depression some economists, in particular Veblen (1919), warned that a crisis was coming. Orthodox economists, including Irving Fisher, disagreed. Even after the Depression began, orthodox economists had little useful policy advice. English economist Arthur Pigou argued that falling wages would reduce unemployment, and Andrew Mellon, a US banker and Treasury Secretary, even welcomed 'liquidation' as the route to recovery. As we will see, mainstream economists also largely failed to warn of the impending GFC.

We will examine the reasons why most mainstream economists were taken by surprise when the GFC hit. In retrospect, it should not be surprising that few mainstream economists foresaw it. Their approach to cycles is based on a fundamental misunderstanding of the dynamics of the capitalist economy. At best, their theory applies to a world in which money is neutral and in which the internal dynamics of the economic system drive it towards equilibrium, as we will see in the next section.

32.2 Why Didn't Mainstream Macroeconomics Foresee the GFC?

After the GFC devastated much of the global economy, the Queen of England pointedly asked her economic advisors why none of them had foreseen the crisis. The irony was that many in the mainstream had in fact been celebrating what US Federal Reserve Chairman Ben Bernanke (2004) labelled "The Great Moderation" – a new golden era of economic stability. They claimed that crises had become much less likely for several reasons.

Central bankers had become more focused on inflation fighting, and had adopted a new strategy of communication with markets that built market confidence in the actions of policymakers. The lower inflation around the world supposedly promoted more stable economic growth. Markets in turn had created a variety of financial innovations that reduced risk and allocated much of the remaining risk to those participants who could handle it. Examples of such innovations included the use of derivatives to hedge risks, such as interest rate risk and exchange rate risk, securitisation that shifted risk out of the banking system, and credit default swap 'insurance'.

Globalised capital markets allowed the diversification of portfolios as well as diversified supply chains which protected corporations from idiosyncratic risks (for example, from earthquakes that might destroy a factory to a striking local labour union). Finally, new mathematical algorithms and lightning-fast computation made it possible to ascertain risk more accurately.

In addition, if all these innovations failed to protect markets from crisis, it was believed that the central bankers would intervene quickly to rescue the markets. The great faith on the part of market participants in the power and wisdom of the central bankers, especially those in charge of the US Federal Reserve, the Bank of England, and the European Central Bank, convinced them that the risk of a crisis was quite small.

Many referred to the 'Greenspan Put', which was the notion that Alan Greenspan, Chairman of the Federal Reserve from 1987 to 2006, had removed most downside risks by coming to the rescue whenever any market stumbled. This presented a 'heads I win, tails you lose' proposition whereby Greenspan would effectively 'socialise any losses and privatise the gains', and encouraged excessive risk taking. There is little doubt that these perceptions promoted euphoric expectations in financial markets.

In the aftermath of the GFC, the US government undertook an investigation to determine the causes of the crash. The Financial Crisis Inquiry Commission (FCIC) was created in May 2009, and released its report in 2011. The Report makes a strong case that the crisis was foreseeable and avoidable. It was not the outcome of a severe, but low probability, random 'shock'. Rather, it was caused by the failure of the biggest global banks to behave correctly even though they were operating under the noses of our 'public stewards' (government banking regulators and supervisors).

According to this remarkable report, the GFC represented a dramatic failure of corporate governance and risk management, which was in large part a result of an unwarranted and unwise focus on trading (effectively, gambling) by banks and the rapid growth of financial balance sheets. At the same time, government policymakers and regulators continually pushed for deregulation and de-supervision in favour of 'self-regulation' and 'self-supervision'. This dangerous combination is what caused the crisis, according to the Report.

While the Report does 'name names' and identify undesirable practices, there is some danger in focusing on the bad individuals' financial practices and events. In the next section, we will put the crisis in the context of the long-term post-war transformation of the financial system, and highlight the structural deficiencies of the whole system.

The crisis *was* foreseen by some market participants, including employees of Goldman Sachs, and some heterodox economists. On the other hand, why did few orthodox economists see it coming? Chairman Alan Greenspan gave a succinct answer when he was grilled in the US Congress. His model of the way the economy worked was wrong (see Box 32.1). Like most orthodox economists, Greenspan had believed that free markets would equilibrate.

He considered that the 'invisible hand' of the market guides rational individuals to maximise their own individual utility, and by doing so they also maximise social welfare. This view considered that it would be irrational for banks to make loans that borrowers could not repay. Investment banks would not bet against their own customers.

Accordingly, market participants would not take on excessive risk, that is, take positions in those assets with a negative risk return trade-off. Further, rational economic agents would not take on excessive leverage, or excessively illiquid positions. While there could be unscrupulous and unwise behaviour by some individuals, it was believed that the market would punish them and eliminate their less successful strategies.

As Chairman Greenspan admitted, this world view was wrong. The 'invisible hand' guided financial institutions to make bad loans, to bet against their customers, to take on excessive risks, and to increase leverage while reducing liquidity to the point that a small uptick of losses would endanger the entire financial system.

BOX 32.1 GREENSPAN ADMITS FLAWED MODEL

REP. HENRY WAXMAN: The question I have for you is, you had an ideology, you had a belief that free, competitive – and this is your statement – "I do have an ideology. My judgment is that free, competitive markets are by far the unrivaled way to organise economies. We've tried regulation. None meaningfully worked." That was your quote.

You had the authority to prevent irresponsible lending practices that led to the subprime mortgage crisis.... And now our whole economy is paying its price.

Do you feel that your ideology pushed you to make decisions that you wish you had not made?

ALAN GREENSPAN: Well, remember that what an ideology is, is a conceptual framework with the way people deal with reality. Everyone has one. You have to – to exist, you need an ideology. The question is whether it is accurate or not.

And what I'm saying to you is, yes, I found a flaw. I don't know how significant or permanent it is, but I've been very distressed by that fact.

REP. HENRY WAXMAN: You found a flaw in the reality...

ALAN GREENSPAN: Flaw in the model that I perceived is the critical functioning structure that defines how the world works, so to speak.

REP. HENRY WAXMAN: In other words, you found that your view of the world, your ideology, was not right, it was not working?

ALAN GREENSPAN: That is – precisely. No, that's precisely the reason I was shocked, because I had been going for 40 years or more with very considerable evidence that it was working exceptionally well.

Source: Excerpted from US Government Publishing Office (2008) House Hearing, 110th Congress – The Financial Crisis and the Role of Federal Regulators

Serial No. 110-209. Available from https://www.govinfo.gov/app/details/CHRG-110hhrg55764/CHRG-110hhrg55764.

Greenspan was a practitioner and policymaker, not an academic economist. He worked in the 'real world' of how economies function, not isolated up in some university ivory tower. Up to the crisis, he was dealing every day with financial markets data and participants. Yet he still got it all wrong because, as he admitted, his free market ideology was wrong. It provided the wrong interpretation of the real world information that he was processing.

Mainstream economists largely adopt the same ideology, but they know much less about real world financial institutions. Much of their theorising does not even permit the existence of anything like a real world financial institution. Many of their models do not have money in them.

REMINDER BOX

In Chapters 29 and 30, we offered a quick review of the main orthodox schools of thought.

- Money is largely reduced to a medium of exchange.
- Financial institutions simply intermediate savings provided by depositors, lending them on to borrowers who want to finance investment in plant and equipment.
- There is no uncertainty about the future, just risk with known probability distributions.
- Those who do take risk are rewarded with a higher return that is just sufficient to compensate for the extra risk.
- The mathematically rigorous models (Dynamic Stochastic General Equilibrium, DSGE) assume that no one ever defaults.

These are the models now used by all of the central banks of the most advanced capitalist economies.

However, despite their failure, mainstream economists and central banks still use economic models that assume that all borrowers always pay all debts as they come due. In such a world, there is no point in assessing credit risk, because all borrowers are equally creditworthy. A crisis like the GFC would be literally impossible. It is unsurprising that policymakers and the academic economists who created these models would fail to see a crisis coming.

After the crisis hit in the US, some mainstream economists claimed that the problem was caused by the US Federal Reserve because it had left interest rates too low for too long, causing the economy to overheat. The reality is that the US Federal Reserve had sharply raised rates, beginning in 2004, even though there was little evidence of inflation or of unemployment falling so low that it might trigger rapidly rising wages. These rate hikes had little immediate effect. Indeed, most of the worst excesses of the real estate and commodity market bubbles in the US occurred after the US Federal Reserve had raised interest rates (see the next section for discussion of the bubbles).

In any event, many mainstream economists, as well as some US Federal Reserve researchers, denied that there was any bubble problem. They argued that, even if there were problems in asset markets, it was not the US Federal Reserve's job to overrule the views of millions of financial market participants. Thus, if those participants were gung ho about real estate and commodity prices, the Board of Governors sitting in Washington, DC had no right to limit their activity. While it is true that some members of that Board secretly worried about excesses in asset markets, the public face of the US Federal Reserve followed the new Chairman's Great Moderation line.

We now know from the transcripts of the Federal Open Market Committee meetings that when the crisis did hit, the US Federal Reserve was surprised and continued to be surprised throughout the following months and years as problems spread from institution to institution, from one financial market to another, and finally across the globe.

As we will see in the next section, the US Federal Reserve eventually had to originate a total of $US29 trillion in loans to prop up the global financial system. This sum is so large as to be nearly incomprehensible and certainly was never contemplated by policymakers or academic economists.

What is most important to notice is that those loans had almost nothing to do with protecting money as the medium of exchange. They were necessary because of the layering of financial debt on financial debt that linked the balance sheets of financial institutions, consisting of both the regulated banks, in addition to the unregulated shadow banks.[1]

While many blame the crisis on liquidity problems, the liquidity crisis was not the outcome of an irrational bank run, but instead reflected an accurate appraisal of financial institution insolvency. That in turn can be attributed to catastrophic reductions of lending standards and to pervasive fraud.

Once financial institutions came to doubt the solvency of one another, they refused to hold one another's liabilities. So instead of a run on banks instigated by depositors, the economy experienced a run by financial institutions on each other. That is why the total lending of the US Federal Reserve was orders of magnitude greater than the sum of all dollar deposits in all banks.

In the 19th century, Bagehot (1873) argued that in a crisis the central bank must lend reserves without limit; that is, the so-called lender of last resort operation. He recognised that once banks are subject to a run only the central bank can stop it, and it does so by lending reserves so that banks can convert their own liabilities to central bank liabilities.

After the development of deposit banking, banks promised to redeem their deposits for either central bank notes or central bank reserves on demand. Again, this meant that the central bank had to lend either notes or reserves on demand to stop the run. If it did not, the run would spread throughout the banking system as holders of private bank notes or private bank deposits demanded cash, that is, central bank notes.

The payments system would seize up if the central bank refused to do so because no bank would accept another bank's liabilities if it feared that it could not get sufficient central bank notes to cover its own liabilities. No bank could survive a run on its liabilities in which depositors or note holders demanded cash for redemption, so the run would spread from bank to bank.

In the 1930s, deposit insurance was created in the USA and the central bank promised to always act as a lender of last resort. This protected the payments system and prevented bank runs. Even despite the severe nature of the GFC, as well as the probable insolvency of many of the biggest banks, there were no runs on insured deposits in the USA.

There was a run on Northern Rock, a UK bank, in September 2007. This was because the Bank of England offered insurance of only 90 per cent of the value of insured deposits. Of course, no depositor wanted to lose ten per cent of their deposits, so when the crisis hit, depositors wisely tried to withdraw deposits. The UK had to increase the insurance to 100 per cent to stop the run. The Rudd government in Australia introduced a deposit guarantee scheme in October 2008, which saved the large four banks from insolvency by ensuring they could continue to access funds in the wholesale markets to cover their liabilities.

However, there were runs on non-deposit liabilities issued by banks and shadow banks, which were not protected by the US Federal Reserve. Thus, the problems were on the fringes of the financial system, on activities and assets that were ignored by mainstream macroeconomists.

Mainstream economists generally took no account of the complexity of the financial system. If money was incorporated into their models, it took a very simple form, either a fiat money issued by government or a deposit money issued as a multiple of reserves of fiat money held for conversion.

Firms in these models are assumed to finance investment either through retained earnings, loans, or sales of equity. From the late 1950s, most orthodox theory held that it makes no difference which of those three alternatives is used. Following from that, it should not matter if firms are indebted, or if they use their own funds to finance their spending. This can be extended just as well to households, so that debt does not really matter.

Recall from above that the most rigorous orthodox models even presume that default never occurs, which is consistent with their view that it makes no difference if it is a household or a firm that is indebted. When economic agents take on debt, it enables them to optimally allocate spending through time so that the receipt of income is not synchronised with desired spending. All that matters is that lifetime spending and income are equal.

Therefore, mainstream economists were not very concerned with rising debt-to-GDP ratios for households, non-financial corporations, and financial institutions. Indeed, they provided demographic reasons as well as other explanations for the trends, thereby explaining rising debt as a rational (and safe) way to smooth spending streams through time. It is thus unsurprising that they did not notice anything to worry about in the lead-up to the crisis.

32.3 Who Did Foresee the GFC and Why?

Introduction

Bezemer (2009) found that economists who followed the orthodox, equilibrium approach did not foresee the crisis. Further, they could not have foreseen it because they either ignored the financial innovations that contributed to creation of the crisis or they saw the innovations as risk-reducing mechanisms which would increase the resilience of the financial system.

Bezemer picked a dozen economists who had predicted the crisis, some as early as the mid- to late-1990s, and had correctly foreseen the nature of the problems that created it. They had abandoned the equilibrium concept and instead adopted a circular flow approach in which stocks and flows are treated consistently (see Chapter 6). These economists emphasised accounting identities rather than equilibrium based on market clearing. They also explicitly modelled the financial system as separate from the 'real' economy. Finally, they allowed for Keynesian uncertainty rather than adopting optimising agents that only have to deal with known probabilities, that is, with risk.

In Chapter 6, we studied accounting identities and the treatment of stocks and flows following the pioneering work of Wynne Godley (1999) and L. Randall Wray (Godley and Wray, 1999). Godley had warned of the unsustainable 'Goldilocks' economy in the 1990s, a situation that had arisen due to the Clinton fiscal surpluses which meant that the non-government sector was deep in deficit. For almost a decade prior to the GFC, the US private domestic sector ran nearly continuous deficits and the private sector debt ratio grew to 125 per cent of GDP. Similarly, in Australia, the Howard government presided over surpluses for most of its time in power (1996–2007), and these were also associated with a rising private debt ratio.

Minsky's financial instability hypothesis

Drawing on the financial instability approach of Hyman Minsky, some economists drew attention to the growing fragility of the financial sector. In the early 1990s, Minsky adopted a 'stages' approach, to describe the long-term transformation of the financial system since the late 19th century (Wray, 2009). He argued that capitalism had gone through three main stages over about one hundred years.

In the early 20th century 'finance capitalism' assumed importance. It was dominated by investment banks that provided finance for corporations. However, by the late 1920s these were mostly financing speculation in financial assets, particularly in equities issued by subsidiary trusts of the investment banks themselves.

These were little more than pyramid schemes.[7] Speculation occurred in essentially worthless shares, much like the infamous schemes of Ponzi or the modern-day Bernie Madoff. Goldman Sachs played a prominent role in the scams.

In Minsky's view, the Great Depression ended the finance capitalism stage and ushered in a much more stable era with the New Deal reforms of the financial sector, plus a much bigger role for the federal government in managing the economy. He called this 'managerial-welfare state' capitalism, in which the 'Big Bank' (US Federal Reserve) and 'Big Government' (Treasury) promoted stable economic growth, high employment, and rising wages and falling income inequality (see Chapter 26).

The US had entered its economic 'golden age', which lasted from the end of the Second World War through the early 1970s. While Minsky's own work focused on the USA, much of his analysis also applies to other major developed capitalist countries, which had their own versions of the American New Deal and experienced robust growth, low unemployment, and mild recessions over this period.

The absence of deep recessions and severe financial crises encouraged innovations that increased financial instability. Further, for reasons that we will not explore, conservative politicians and economists were able to slowly chip away at the New Deal reforms that promoted growth while providing social protection.

After 1974, US median male earnings stalled and began to fall as workers lost effective representation by unions and the social safety net was undermined. Deliberately increasing unemployment was a tool used by policymakers to keep inflation low.

There were several important pieces of legislation to reduce regulation, including the Gramm-Leach-Bliley Act of 1999 that ended the Glass-Steagall separation of investment banking from commercial banking. There was also a much greater reliance on self-supervision by financial institutions. They were able to capture a greater share of national profits and this enhanced their political power, making it possible to further subvert or eliminate regulations so that they could gain an even larger share of profits.

This transformation was referred to as 'casino capitalism', and 'financialisation'. It was like 'finance capitalism', with shadow banks now playing the role played by investment banks in the 1920s. In the early 1990s, Minsky was warning that this newest phase, which he called 'Money Manager Capitalism', was highly unstable and prone to deepening crises.

An important characteristic of this stage was the decline in importance of the regulated banking sector and the rise of 'managed money' in the 'shadow banking sector', as shown in Table 32.1. This provided a justification for dropping the New Deal reforms so that the banks could 'compete' with the new intruders who were poaching business.

Managed money includes pension funds, retirement savings accounts, university endowments and sovereign wealth funds. Like banks, these entities issue liabilities (for example, pension funds have liabilities to pensioners) and purchase assets such as the liabilities of banks and shadow banks, government bonds, and the mortgage debt of homeowners.

Hence, while private sector debt was growing rapidly, an increasing share was being held by the lightly regulated and supervised shadow banking system. On the other hand, the commercial banks and thrifts (savings institutions) were not only closely supervised by government agencies, but their liabilities were largely government guaranteed. The liabilities issued by managed money were not government guaranteed, and thus were risky.

Some orthodox economists thought that this regulatory environment was appropriate. Potential losses for government under the guarantee were limited due to banks being protected from high-risk assets, while there were incentives for those who held liabilities of shadow banks to discipline their behaviour.

However, these arguments turned out to be false. When the GFC hit, governments bailed out some shadow banks, such as AIG, and extended deposit insurance to some. Governments also bought or lent against risky assets (such as asset backed securities). We now know that no one was disciplining the shadow banks which bought assets that they expected to fail. Also, the regulated commercial banks were financially connected to the

Table 32.1 Share of financial institutions, 1945–2008, percentage of total financial sector assets

	1945	1950	1960	1970	1980	1990	2000	2008
Commercial banking total financial assets	59.8	50.3	36.9	36.4	33.5	25.3	19.1	24.1
Savings institutions	10.7	13.2	18.0	17.8	17.9	10.0	3.5	2.6
Credit unions	0.1	0.3	1.0	1.2	1.5	1.6	1.3	1.4
Property casualty insurance companies	2.6	4.0	4.2	3.6	4.1	4.0	2.4	2.3
Life insurance companies	18.3	21.0	18.6	14.2	10.5	10.3	8.9	7.8
Private pension funds	1.6	2.0	6.6	8.7	11.6	12.4	12.7	7.9
State and local government employee retirement funds	1.1	1.6	3.2	4.2	4.4	5.5	6.5	4.0
Money market mutual funds	0.0	0.0	0.0	0.0	1.7	3.7	5.2	6.5
Mutual funds	0.5	1.1	2.7	3.3	1.4	4.6	12.6	9.4
Closed end funds	0.4	0.7	1.0	0.4	0.2	0.4	0.4	0.4
Government sponsored enterprises	1.0	1.1	1.9	3.3	4.4	3.6	5.6	5.8
Agency and GSE backed mortgage pools	0.0	0.0	0.0	0.3	2.6	7.7	7.1	8.6
Issuers of asset backed securities	0.0	0.0	0.0	0.0	0.0	2.0	4.3	7.1
Finance companies	1.8	3.2	4.7	5.0	4.8	4.5	3.5	3.2
Real estate investment trusts	0.0	0.0	0.0	0.3	0.1	0.2	0.2	0.4
Security brokers and dealers	2.1	1.4	1.1	1.1	1.0	2.0	3.5	3.8
Funding corporations	0.0	0.0	0.0	0.1	0.4	1.9	3.3	4.7
Managed money	3.6	5.4	13.5	16.7	19.3	26.7	37.4	28.2

Source: Data from US Treasury, Federal Reserve Flow of Funds Accounts.

shadow banks in numerous ways so that they had to absorb a high share of the losses due to risky behaviour by money managers.

Many commentators referred to the advent of the GFC as a 'Minsky Crisis' or 'Minsky Moment', invoking his famous 'financial instability hypothesis' (FIH) that described the transformation of an economy from a 'robust' to a 'fragile' financial structure:

- A run of good times would encourage ever greater risk taking by investors, financial institutions, and entrepreneurs.
- They tended to borrow more relative to expected earnings, and innovate with riskier financial instruments.
- In addition, regulators might relax rules in the belief that downside risks had been reduced.

Minsky's description seems to perfectly describe the last few decades of US experience, during which financial crises became more frequent and increasingly severe. We could list, for example, the Savings and Loan crisis of the 1980s, the stock market crash of 1987, the Developing Country debt crises (1980s, early 1990s), the Long-term Capital Markets and Enron fiascos, and the dotcom collapse.

Each of these crises led to US government intervention that prevented a major downward spiral of financial markets or of the economy, although in some cases, temporary recessions followed the crises. Indeed, after the dotcom crisis, the belief took hold that a new Great Moderation made major downturns impossible. All of this encouraged more risk taking, more financial layering and leveraging (debt issued against debt, with little net worth backing it up). All that behaviour was consistent with Minsky's FIH.

Many if not most of the new practices served no social purpose beyond rewarding the top management of financial institutions. At the same time, the structure of incentives and rewards was changed such that risky bets, high leverage ratios, and short-term profits were promoted over long-term firm survival and returns to investors.

In the money manager capitalism stage, every money manager had to beat the average return to retain clients, which is of course statistically impossible. However, with such incentives and with virtually no government regulation or oversight, this encouraged not only risky behaviour but also actions which were ethically compromised.

In Minsky's view, the rise of these managed funds was facilitated by the success of the earlier managerial welfare state capitalism: the absence of depressions and relatively good growth, along with policies that favoured private pensions and allowed financial wealth to grow over the entire post-war period. Although financial crises came along and wiped out some wealth, each crisis was sufficiently contained that most wealth survived, and its growth quickly resumed.

There was a 'Gresham's law' ('bad money drives out good') in operation, with those institutions that could quickly reduce capital ratios and loss reserves being able to increase net earnings and thus rewards to management and investors. Further, there was a shift to the maximisation of share prices as one of the main goals of management, which supposedly aligned the interests of shareholders and top managers who received stock options as part of their compensation. That in turn encouraged a short-term focus on performance in equity markets, which was accomplished through market manipulation (both legal and illegal).

Significantly, the sheer volume of financial wealth under management outstripped socially useful investments. To keep returns high, money managers and bankers had to turn to increasingly esoteric financial speculation in areas that not only did not serve the public purpose, but actively subverted it.

An example would be the rise of index speculation in commodities markets that drove up global prices of energy and food, leading to hunger and even starvation around the world. The dotcom bubble is another example. Speculators drove up the prices of stocks of internet companies with no business model or prospective profits, and the inevitable crash wiped out hundreds of billions of dollars of wealth.

Yet another example is the US real estate boom that began before 2000 and finally collapsed in 2007, triggering the GFC. It was the biggest speculative boom in US history, and was driven by money managers who created complex securities and derivatives for speculative bets. Nothing is more symptomatic of the speculative excess than the special collateralised debt obligations (CDOs) created by investment banks to allow hedge funds to bet against homeowners and the holders of securitised mortgages.

The CDOs would pay off if the homeowners were not able to make their mortgage payments, meaning that the securities the investors bought would fall in value. The investors would lose, the homeowners would lose their homes, and the hedge fund would win! The investment bank won too, because it got a fee income for creating the securities and for creating the CDO instrument that let the hedge fund win when the homeowner and the investor in the homeowner's mortgage lost. It was this demand by investment banks and other speculators for risky instruments on which they could place bets that generated the risky 'subprime' and 'Alt A' mortgages that eventually brought on the GFC.

The rise in inequality

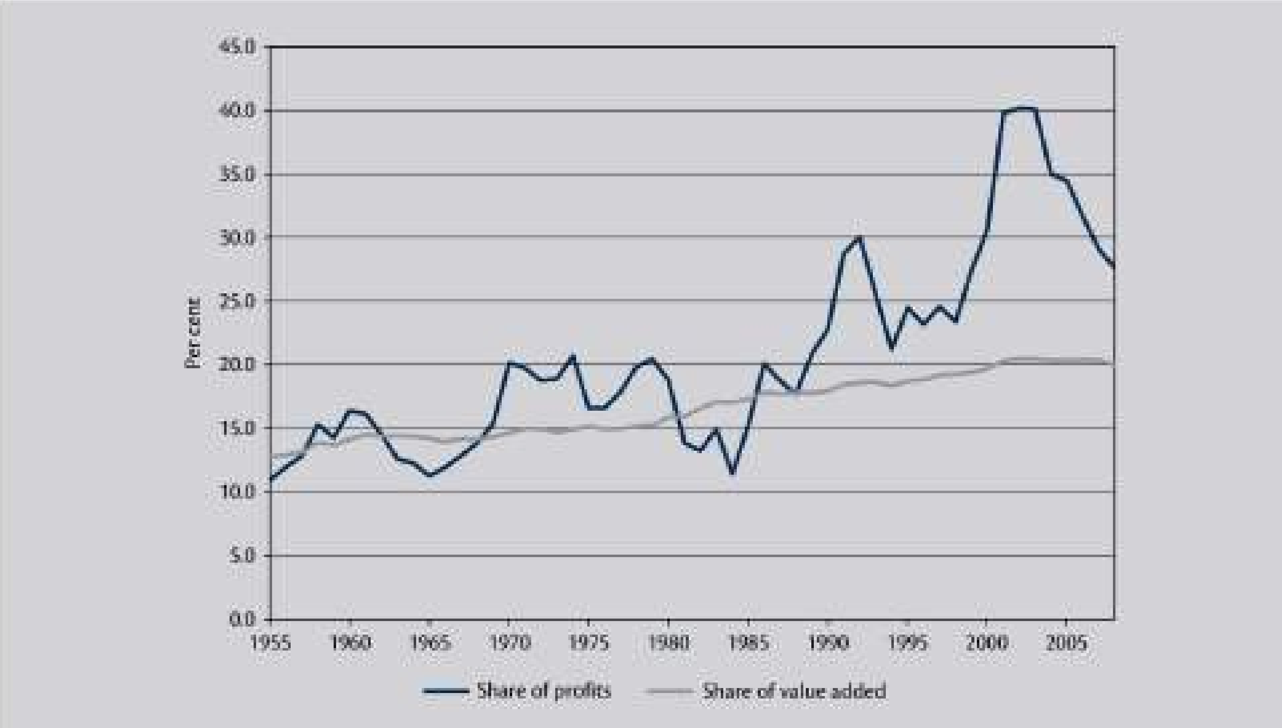
Minsky and Whalen (1996) also linked rising economic insecurity to the money manager phase (see also Minsky, 1996). Inequality began to rise, along with trend increases to the unemployment rate, at least until the boom and bust cycle that began with President Clinton. After a decade (1996–2006) with some improvement in the macroeconomy, and unemployment trending somewhat lower, economic growth improving, and the increase in poverty ceasing, the GFC caused a significant increase in unemployment along with rising poverty, and inequality rose to record levels.

Figure 32.1 shows the financial sector's share of corporate profits and value added from 1955 to 2008. The financial sector grew relative to the non-financial sectors. When the GFC commenced, the financial sector accounted for 20 per cent of US national value added and 40 per cent of corporate profits. It was an autonomous source of growth and of rising inequality due to high compensation in the sector.

Up to half of the college graduates from the elite colleges went into the financial sector because rewards there were far higher than in other sectors. Compensation in the top positions quite simply exploded.

As shown earlier, by 2007 the US ratio of debt to GDP reached an all-time peak of 500 per cent. While much discussion in recent years has been about the government debt ratio, the debt of the household sector, as well as

Figure 32.1 Financialisation of the US economy



Source: US Bureau of Economic Analysis.

Note: The data refer to the FIRE (finance, insurance, real estate, rental, and leasing) sector.

those of non-financial and financial businesses, were all much higher as a percentage of GDP. Non-financial business debt was not a huge problem because much of this was due to long-term finance of capital equipment and, after 2000, borrowing by US non-financial businesses was relatively low.

However household debt was a huge problem, and it is clear that in the aftermath of the GFC, household debt remains a serious on-going issue, preventing full economic recovery. Also, there was the unprecedented rise of financial sector indebtedness, which reached almost 125 per cent of GDP and has been largely ignored.

Another aspect of the rise of MMC has been the increase in income and wealth inequality (see Box 32.2). After the Second World War, robust growth plus New Deal institutions and other factors such as the civil rights movement and stronger bargaining strength of labour unions had reduced inequality in the US. Other wealthy capitalist countries saw similar outcomes, and developing nations on average enjoyed fast economic growth and rising living standards over this period. In the US, all of this began to change after the early 1970s for a variety of reasons, including the erosion of New Deal safety nets, diminished union power (in part due to greater competition from imports), and oil price shocks.

Rather than pursuing full employment, many governments (including that of the US) began to see higher unemployment as a safety valve against the higher rate of inflation, which also contributed to lower bargaining power for workers. In addition, economists pushed the notion of a 'government budget constraint' similar to that faced by households, and argued that fiscal discipline was required to ensure government solvency.

However, it is also important to recognise that the rising importance of finance tended to shift income and wealth gains toward the financial sector. Interest income, capital gains, fees charged for financial services, and rewards for speculation in financial assets boosted the share of income going to non-wage sources. As the wage share declined and real wages stagnated in the US, American workers resorted to borrowing to maintain living standards. Together, these developments reversed the post-war trends so that inequality of income and wealth began to climb.

BOX 32.2 MONEY MANAGER CAPITALISM (MMC) AND THE RISE OF INEQUALITY

The surprising outcome over the post-war period is that during expansions the share of income growth that goes to the richest people rises inexorably. Figure 32.2 reproduces a graph created by Pavlina Tcherneva that divides the distribution of income growth during expansions between the bottom 90 per cent and top 10 per cent of the US population.

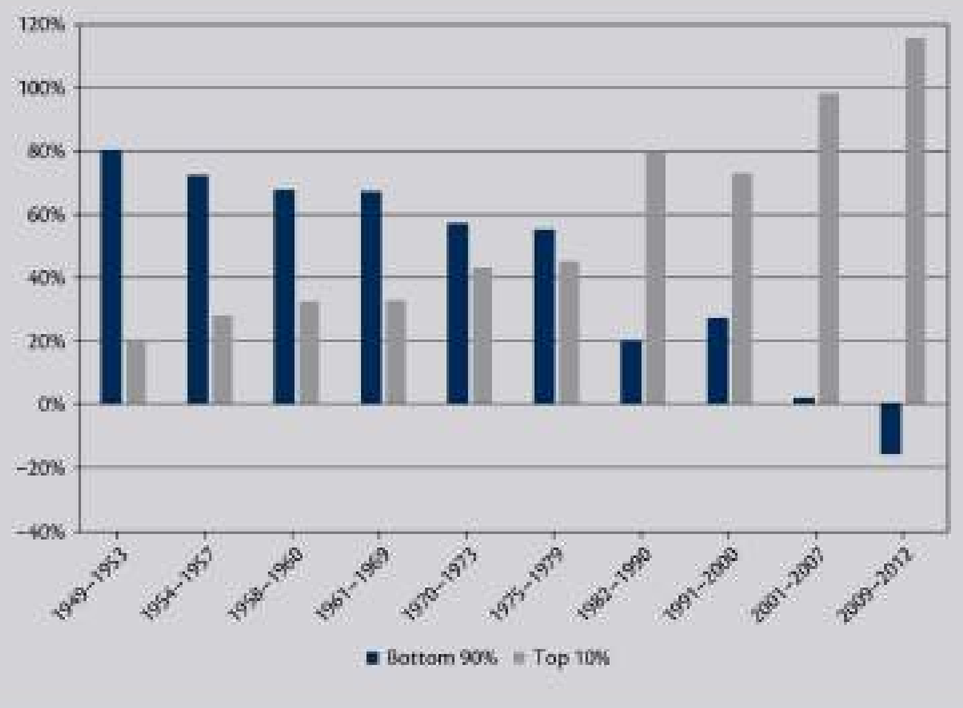
In the first post-war expansion, fully 80 per cent of the gains went to the bottom 90 per cent of the population. That still means that the top 10 per cent got 20 per cent of the gains, which is more than their 'fair share'.

However, the share of the gains obtained by the top 10 per cent rises sharply over the post-war period (with the exception of a slight decline between the expansions of the 1980s and 1990s). In the 'recovery' of 2009–12 after the GFC, the income of the top 10 per cent grew by more than 100 per cent. Thus, the bottom 90 per cent was actually worse off after four years of 'recovery' than they had been at the end of the GFC's recession in 2009.

The view held by 'Keynesians' during the 1960s was that 'the rising tide lifts all boats', so that economic growth would benefit everyone. Figure 32.2 demonstrates that until the late 1970s that while growth raised the boats of those at the bottom at least a little, it seems to have raised the yachts of those at the top more than proportionately. Further, since the 1980s, the gains from growth have disproportionately favoured the top 10 per cent to an ever-increasing degree.

From these data, one could hypothesize that growth alone does not necessarily raise the boats of those at the bottom, and that the transformation of the structure of the economy toward MMC seems to have reduced the benefits of growth. Economic growth now accompanies rising inequality. While we cannot conclude that growth, *per se*, *causes* inequality, it does not appear to ameliorate it.

Figure 32.2 Distribution of average income growth during expansions, US, 1949–53 to 2009–12



Source: Tcherneva, P.R. (2017) Analysis of Piketty/Saez data and NBER. Reproduced by permission of the Levy Economics Institute of Bard College

Note: Includes capital gains.

Neoclassical economists explain rising inequality in terms of the market rewarding the portion of the population that is increasing its productivity ('skill-biased technical change'). The most common explanation for rising inequality is that the growth of the 'knowledge economy' has increased the demand for those with higher education, particularly for workers in the STEM (Science, Technology, Engineering and Math) sector.

At the same time, the demand for those with less education has fallen, either because the jobs can be done much more cheaply in the developing world, or because automation replaced many of them. The market is sending out signals to young people that they should change their career goals and prepare for these new realities.

Those that do not make these changes apparently are simply expressing their preferences to work in lower-paid occupations, perhaps because they enjoy them. Alternatively, they may recognise that they do not have the capacity to upgrade their skills to those needed in the STEM sector.

With this 'free market' frame of reference, it is not surprising that mainstream economists either did not notice the significant rise of inequality, or if they did, dismissed it as a problem unworthy of policy analysis. Only since the GFC have some mainstream economists, notably Larry Summers and Paul Krugman, begun to worry about the problems posed by rising inequality. Most importantly, they argue that because high-income individuals' marginal propensity to spend is lower than that of others, this shift of income distribution will open demand gaps.

Setting aside the theoretical issues of inequality momentarily to consider a real world situation, Box 32.3 describes the consequences of the US Government's bailout of Wall Street following the GFC.

Further, technological changes, as well as the ageing of the population that have resulted in lower labour force participation rates, also tend to depress demand because those out of the labour force generally have less income. For these reasons, some economists worry that developed capitalist economies might have entered a new era of stagnation in which growth will be chronically constrained. They argue that this justifies the greater use of fiscal policy to reduce the demand gap.

While such mainstream arguments are not entirely wrong, they do not recognise the scope of the problem. Just as they could not see the GFC coming, they cannot see the fundamental problems facing our economies. Thus, they cannot formulate a comprehensive set of policies to resolve them. Before we turn to the topic of designing macroeconomic theory and policy for the future, let us review another problem area that some economists got right, but most economists never saw coming, namely the problems in the Eurozone.

BOX 32.3 HOW THE US GOVERNMENT'S BAILOUT OF WALL STREET PROMOTED INEQUALITY AND REWARDED BAD BEHAVIOUR

The US Federal Reserve's bailouts of Wall Street stretched, and might have violated, both the law as established in the Federal Reserve Act (and its amendments) and well-established procedure. There is a long tradition accepted by central bankers of a distinction between continuous versus emergency borrowing by banks at the central bank's 'discount window'. We look at the case of the US Federal Reserve's rescue of the US and global financial systems.

Briefly, the US Federal Reserve is permitted to lend (freely as Bagehot recommended) to resolve a liquidity crisis, but it has long refused to provide 'continuous' lending. Here the idea is that the US Federal Reserve should stop a liquidity crisis, but then solvent financial institutions should quickly return to market funding of their positions in assets. The GFC started in 2008. Four years later the US Federal Reserve was still lending, and at 'subsidised' (below-market) interest rates. That clearly goes beyond emergency short-term lending in a crisis.

The US Federal Reserve is also generally prohibited from lending to 'non-bank' financial institutions (what we now call 'shadow banks') that are not members of the Federal Reserve System and do not issue FDIC-insured deposits. However, an exception is granted in the Federal Reserve Act's '13(3)' provisions, which allows the US Federal Reserve to lend in 'unusual and exigent' conditions. Certainly, the GFC meets these conditions.

However, the 13(3) restrictions are tight, and the US Federal Reserve seems to have stretched the law. Some might object that while there was some questionable, possibly illegal activity by the nation's central bank, was it not justified by the circumstances?

The problem is that this 'bailout' validated the questionable, risky, and in some cases illegal activities of top management on Wall Street – the people who ran the 'control frauds' in the terminology of white collar crime expert William Black (2005).

Most researchers agree that the effect of the bailout has been to continue, if not increase, the distribution of income and wealth flowing to the top one-tenth of one percent. It has kept the same management in control of the worst serial abusers, namely Goldman, Bank of America, Citigroup, and JPMorgan Chase, which paid record bonuses to top management.

Some of the fraudulent activity has been exposed, and the top banks have paid numerous fines for bad behaviour. Yet, Washington has been seemingly paralysed with Eric Holder, US Attorney General during the Obama Presidency, not even beginning a single investigation of criminal behaviour by top management.

What should have been done? Bagehot's (1873) recommendations for a central bank to 'lend without limit' at 'penalty interest rates' and against 'good collateral' are sound, but must be amended. Any of the 'too big to fail' financial institutions (what Black (2005) calls 'systemically dangerous institutions') that needed funding should have been required to submit to US Federal Reserve oversight.

Top management should have been required to submit resignations as a condition of lending, with the US Federal Reserve or Treasury holding the letters until they could decide which should be accepted. The Attorney General's office should have been called in to investigate all top management, to prosecute crimes, and to pursue gaol time for those who were convicted.

Short-term lending against the best collateral should have been provided, at penalty rates. A comprehensive 'cease and desist' order should have been enforced to stop all trading, all lending, all asset sales, and all bonus payments. The Federal Deposit Insurance Corporation should have been called in (in the case of institutions with insured deposits), but in any case, the institutions should have been dissolved according to existing law: at least cost to the Treasury and to avoid increasing concentration in the financial sector.

These actions would have left the financial system healthier and smaller; they would have avoided the moral hazard problem that has grown over the past three decades as each risky innovation was validated by a government-engineered rescue; and they would have reduced the influence that a handful of huge banks have over policymakers in Washington.

32.4 Lessons That Can be Learned About Sovereign Currency From the Eurozone Crisis

European integration was a grand plan which was probably driven by the best of intentions. However, a number of Modern Monetary Theory (MMT) advocates argued from the outset that the design of the European Monetary Union (EMU) was fatally flawed. At the very least, the euro currency was adopted before fiscal integration was achieved under an authority with sufficient sovereignty to protect the member nations. Mitchell (2015) provides a detailed account of these failures in the design of the system.

There was no treasury which operated with a sovereign currency that was answerable to the European Parliament, so its spending required contributions from non-sovereign member states. Indeed, the apparent belief was that permanent fiscal austerity was the appropriate strategy for growth, which meant that all EMU members were forced to adopt austerity measures.

The Maastricht Criteria – and later, the Fiscal Compact – provided the target outcomes for each EMU member, which were associated with the implementation of austerity measures to address the so-called fiscal crisis, and the sanctions for non-compliance.

During the first decade of the Eurozone, many critics focused on the policy of the European Central Bank (ECB), arguing that its monetary policy was too tight. Others argued that the Maastricht Criteria were too tight. While both criticisms had some validity, they missed the main problem: Italy and other countries had become the equivalent of a State such as Louisiana or Mississippi in the United States or NSW or Victoria in Australia, which do not issue their own currencies. The difference, however, between these States and the Member States of the Eurozone is that the latter do not enjoy the benefit of a federal government that can make monetary transfers to states in crisis.

Given the EMU structure, individual Eurozone members faced two problems:

- When there was a deep recession, their fiscal positions would automatically move into deep deficit. The problem would not be the Maastricht Criteria *per se*, since almost all euro nations persistently violated those criteria over the decade before the crisis which began in 2009, but rather that markets would raise risk premia on their debt, which occurred in 2008. This caused interest rates to rise sharply in a manner that reinforced the rise in deficits in a vicious cycle. With no fiscal authority to come to their rescue, Eurozone countries had to rely on the charity of the ECB to keep their interest rates down. With the ECB operating under the thumb of the German Bundesbank, that was always unlikely.
- Individual nations were responsible for their own banking systems, but they did not have the fiscal capacity to bail them out. Again, there was no equivalent to a fiscal authority in Brussels to assist the national governments that were burdened by debt run up by private banks, debt that could easily be more than prevailing government spending or even GDP.[3]

One of the goals of integration was to free up labour and capital flows by removing barriers so that factors of production could cross borders. What is important for our discussion is that this also enabled Eurozone banks to buy assets and issue liabilities all over the monetary union.

Further, the deregulation and de-supervision of banking outlined in the Basel Accords allowed banks to undertake the same sort of risky schemes that Wall Street's banks had pursued. The result was that by 2007 the EU had built up a real estate bubble that was almost identical in size to the US's bubble, and again this was largely unrecognised by mainstream economists who held faith in the wisdom of 'free markets'.

The Irish banks had ramped up their lending across Europe, growing their liabilities to multiples of Irish GDP. Then, when their bets went bad, the Irish government had to bail them out, boosting fiscal deficits and raising government debt to uncharted territory. Again, this was a design feature of the EMU and the EU more generally, so it was not just euro banks that engaged in excessive lending. This also occurred in Iceland and the UK. It is probable that this was not foreseen, for the reasons discussed above. Mainstream economists had little understanding of financial markets because their theory and models had no room for the real world complexity of modern finance.

Also important to the continuing crisis in the Eurozone, which actually sank deeper into crisis as the US and UK recovered, was the ability of bank depositors to costlessly shift euro deposits from one bank to another, anywhere in the EMU. This is possible due to the so-called Target 2 facility. For example, any depositor of a Spanish bank can move deposits to a German bank. Such a shift requires that the central bank of Spain obtain reserves that are then credited to the central bank of Germany. If deposits tend to flow from the periphery nations, their central banks go ever more deeply into debt to the ECB to obtain reserves that accumulate in the accounts of the strongest nations' central banks, such as the German Bundesbank.

As the crisis spread among periphery nations, the 'Troika' (consisting of the IMF, the ECB and the European Commission) demanded austerity as a condition for financial assistance. Yet the only way any individual state could grow in the face of austerity would be by operating a 'beggar thy neighbour' mercantilist policy to suck demand out of the other states. This was done by achieving a positive trade balance, that is, with exports exceeding imports.

Exporting earns reserve credits for the nation's central bank, and creates income and jobs for the domestic economy. Germany excelled at that. So, while the whole idea behind unification was to prevent the type of unneighbourly behaviour that had led to two World Wars within Europe, the construction of the EMU was guaranteed to promote it. The EMU rewarded self-interested behaviour by any member willing to pursue it, and Germany reaped most of the rewards.

In addition, financial institutions had the incentive to take on more risk, which in addition to free 'capital' movements provided the Eurozone with the preconditions for a financial crisis. However individual nations remained fully responsible for their own outsized financial institutions should a crisis hit.

Without fiscal sovereignty, the first serious financial crisis would undermine the fiscal position of some member nations. That happened in Ireland. Others followed, falling like dominoes.

It is wrong to point the finger at Troika-imposed austerity for the initial problems on the periphery. Permanent austerity was always the plan. This is not new. It is the way that non-sovereign governments must operate in the absence of a sovereign fiscal authority. The problem was never one of profligate Mediterranean countries with lax fiscal policy. Given the rules, no euro nation should ever have run chronic deficits of any size; and so none should have run up any significant debt ratio. By design, these are not sovereign countries in the currency sense, since they abandoned their own sovereign currencies years ago in favour of a foreign currency.

As Wynne Godley and L. Randall Wray warned in 1999, the member nations would become similar to colonies that used the currency of their colonisers. Further, like any nation that gives up its sovereign currency, each one lost the ability to run ongoing fiscal deficits, which is a requirement for all the nations that lose the beggar-thy-neighbour race for trade surpluses.

The EMU has been saved to date (2018) because the ECB introduced a large quantitative easing programme, which effectively amounted to funding the fiscal deficits of the Member States in contravention of the spirit of the Treaty of Lisbon's 'no bailout clause'.

However, the monetary union cannot work effectively without fiscal authority while relying on a reluctant central bank to go against its programming. Hence, it is not at all clear that the EMU will be saved in the long term.

Why didn't mainstream economists see this coming? Because they believed that monetary policy should be independent of fiscal policy. They believed that governments should face a tight 'government budget constraint' enforced by market discipline. Markets *should* increase interest rates to punish irresponsible government deficits.

Mainstream economists particularly oppose 'money financed' fiscal deficits that are supposed to result in inflation. These economists do not understand sectoral balances, specifically how a fiscal constraint impacts on the non-government balance sheet. Equally, they do not recognise the impact of external balances on domestic balances. Also, they believe that 'market discipline' can keep financial institutions in line, without much regulation and supervision by government authorities. For all these reasons, they believed that the design of the Eurozone was fit for purpose.

It is no surprise that they did not anticipate either the GFC or the associated Eurozone crisis.

Conclusion

The conceptual, theoretical and policy deficiencies of mainstream macroeconomics have been highlighted in the period following the GFC and the ongoing economic crisis in the European Monetary Union.

Modern Monetary Theory has provided a rigorous conceptual framework within which to understand the adverse economic outcomes. This has attracted some attention amongst leading mainstream thinkers including Paul Krugman, but also hostility from a range of academic economists attached to different schools of thought, which suggests that MMT is gaining some traction within the discipline.

References

Bagehot, W. (1873) *Lombard Street: A Description of the Money Market*, New York: Scribner, Armstrong.

Bernanke, B. (2004) "The Great Moderation", Board of Governors of the US Federal Reserve System, at the meeting of the Eastern Economic Association, Washington, DC, 20 February.

Bezemer, D.J. (2009) "No One Saw this Coming. Understanding Financial Crisis Through Accounting Models", Research Report, Research Institute SOM (Systems, Organisations and Management), the University of Groningen, Netherlands.

Black, W. (2005) *The Best Way to Rob a Bank is to Own One: How Corporate Executives and Politicians Looted the S&L Industry*, Austin, TX: University of Texas Press.

FCIC (Financial Crisis Inquiry Commission) (2011) *Financial Crisis Inquiry Report: Final Report of the National Commission on the Causes of the Financial and Economic Crisis in the United States.*

Godley, W. (1999) "Seven Unsustainable Processes: Medium-term Prospects and Policies for the United States and the World", *Strategic Analysis*, Levy Economics Institute of Bard College, January. Available at: http://www.levyinstitute.org/publications/seven-unsustainable-processes, accessed 10 July 2017.

Godley, W. and Wray, L.R. (1999) "Can Goldilocks Survive?", *Policy Note, 1999/4*, Levy Economics Institute of Bard College, April. Available at: http://www.levyinstitute.org/publications/can-goldilocks-survive, accessed 10 July 2017.

Mitchell, W.F. (2015) *Eurozone Dystopia: Groupthink and Denial on a Grand Scale*, Cheltenham: Edward Elgar.

Minsky, H.P. (1996) "Uncertainty and the Institutional Structure of Capitalist Economies", *Working Paper No. 155*, Levy Economics Institute of Bard College, April.

Minsky, H.P. and Whalen, C.J. (1996) "Economic Insecurity and the Institutional Prerequisites for Successful Capitalism", *Working Paper No. 165*, Levy Economics Institute of Bard College, May. Available at: http://www.levyinstitute.org/pubs/wp165.pdf, accessed 10 July 2017.

Tcherneva, P.R. (2017) "Inequality Update: Who Gains When Income Grows?", Policy Note 2017/1, Levy Economics Institute of Bard College, April. Available at: http://www.levyinstitute.org/publications/inequality-update-who-gains-when-income-grows, accessed 10 July 2018.

Veblen, T. (1919) "On the Nature and Uses of Sabotage", *Dial*, LXVI, April 5, 341–6.

US Government Publishing Office (2008), *House Hearing, 110th Congress – The Financial Crisis And The Role Of Federal Regulators, Serial No. 110-209.* Available at: https://www.govinfo.gov/app/details/CHRG-110hhrg55764/CHRG-110hhrg55764, accessed 8 August 2018.

Wray, L.R. (2009) "The Rise and Fall of Money Manager Capitalism: A Minskian Approach", *Cambridge Journal of Economics*, 33(4), 807–28.

Endnotes

1. A shadow bank is an unregulated financial institution which can engage in credit creation.
2. A pyramid scheme is a business model in which members are recruited and are promised remuneration from their enrolment of others into the scheme, rather than through supplying goods or services.
3. It could be argued that if member countries were at different stages of the business cycle, the imposition of a uniform target interest rate across the Eurozone would inevitably be counter-productive. However within a country, economic conditions in rural areas may be markedly different from those in major urban areas, so the issues associated with the conduct of the 'one size fits all' monetary policy within the Eurozone can apply equally to a single country.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

CHAPTER

33 MACROECONOMICS FOR THE FUTURE

Chapter Outline

Learning Objectives

- Recognise the importance of developing macroeconomic models that are stock-flow consistent.
- Understand the economic capacity of a government which issues its own fiat currency.
- Acknowledge the need for the central bank to be accommodative with respect to liquidity management.
- Understand that the central banks typically do not control the volume of bank lending.
- Appreciate the importance of the choice of exchange rate regime in determining the extent to which a government can exercise discretion in policymaking.

33.1 Introduction

This book has been designed to expose students to *Modern Monetary Theory* (MMT) and apply its principles to contemporary macroeconomic problems. This conceptual and analytical framework is grounded in the principles of functional finance and the theory of money known as Chartalism. When applied to the analysis of a particular macro economy, it is informed by prevailing institutional practices. Typically these practices incorporate constraints on the behaviour of the central bank and/or treasury, restrictions that political and ideological processes have imposed and are often misconstrued as meaningful (involuntary) constraints. MMT removes this veil of ideology to expose the intrinsic features of a modern monetary economy. Indeed, we have tried to avoid focusing on the institutional arrangements associated with any one particular country so as to give the book wide application and relevance.

In this chapter, we summarise some of the key propositions arising from MMT that have been developed in this textbook and which should assist students to revise their understanding of the operation of a modern, developed, capitalist economy. We should emphasise that these insights are not all unique to MMT and draw on a broad heterodox literature.

The Global Financial Crisis (GFC) – and its aftermath – have played a key role in consolidating and refining the MMT's theoretical and policy framework. Also, the extreme macroeconomic conditions experienced in some countries (and resultant data generated) have enabled empirical MMT propositions to be rigorously tested against mainstream alternatives. The evidence is clear. The mainstream predictions of accelerating inflation (from the quantitative easing), rising interest rates (from the fiscal deficits), and government insolvency (from the rising debt) have not been realised. The major MMT propositions have been vindicated by our real-world experience.

Mainstream thinking has also been subject to critical scrutiny – not only by heterodox economists but also by some of its own adherents – in the light of the impact of the GFC and the subsequent policy responses by governments and other institutions, particularly in the European Monetary Union (EMU).

Among mainstream theories and claims that should no longer be taken seriously by policymakers, we can include: Rational Expectations and continuous market clearing; New Classical and Real Business Cycle approaches; neutral money; the New Monetary Consensus, the Taylor Rule and the Great Moderation; the Efficient Markets Hypothesis; Ricardian Equivalence and other versions of the policy irrelevance doctrine. More broadly, arguments made by advocates for deregulation and self-regulation need to be carefully scrutinised.

None of these ideas should be taught as serious economic theory, nor should they form the basis of policy. They are no more relevant to economic theory and policy than are bloodletting techniques to the study of medicine. They are still important, of course, in the study of the history of economic thought as examples of where theorising has occurred without being underpinned by either a coherent set of macroeconomic principles or a clear understanding of prevailing institutional practices, such as how central banks and treasuries actually operate.

33.2 Modelling Framework

The analytical framework outlined in this textbook is disciplined by the use of balance sheets which ensure that modelling is stock-flow consistent. Financial assets equal financial liabilities, and total assets (financial assets plus real assets) equal financial liabilities plus net worth. The first of these is an identity which holds at the aggregate level; the second is true at the aggregate level as well as at the level of every economic entity.

Stock-flow consistency means that every flow must come from somewhere and go somewhere, with flows accumulating to stocks. For example, an income flow can finance a spending flow plus a flow of saving. That spending must be received somewhere as income, and the saving flow must accumulate as a stock of assets.

These flows are linked in turn through the identity based on macroeconomic sectoral balances. National income equals national expenditure and thus GDP, which is the sum of consumption, investment, government spending, and net exports. Most economists accept that at the aggregate level, causation goes from expenditure to income. The reasoning is that a nation can decide to spend more, thereby generating more income, but it cannot simply decide to have more income. The key is that financial institutions can provide finance for the additional expenditure, which then creates more income. While we accept that at the level of the individual household, firm or even local government, income largely determines spending; at the aggregate level it is spending that determines income.

By construction we can then obtain the identity that injections equal leakages: $I + G + X = S + T + M$ (I is investment, G is government spending, X is exports, S is saving, T is tax revenues, and M is imports), which can be rearranged as $(I - S) + (G - T) + (X - M) = 0$. Each set of parentheses encloses a net injection (positive) or a net leakage (negative), the sum of which must be zero.

This balance, derived from the National Income and Product Accounts (NIPA), is similar to the sectoral balance identity of Wynne Godley, which he derived from the flow of funds data: domestic private balance +

government sector balance + foreign sector balance = zero. There are some relatively minor differences between the NIPA and flow of funds versions of this macro identity. The most important one for our purposes is that the NIPA version uses net exports, while the flow of funds version uses a broader measure – the current account balance.

While the NIPA identity is, by definition, always satisfied, it provides powerful insights. For any sector to run a surplus (a positive balance, with its income exceeding its expenditure, or with its leakages exceeding its injections), there must be at least one other sector running a deficit (spending greater than income, or injections greater than leakages).

If the external balance is zero, then a domestic private surplus is equal by identity to the domestic government sector's deficit; and the government's deficit means that it will be issuing debt which is accumulated as net financial saving by the private sector.

On the other hand, for the government to run a surplus (given a zero balance in the external accounts) requires that the private sector runs a deficit, thereby issuing debt accumulated by the government. If the government is running a surplus, then the private sector deficit will be even larger if the external balance is in deficit; in that case foreigners will also accumulate debt issued by the domestic private sector.

For the global economy taken as a whole, the external deficits and surpluses must net to zero. Hence, the global private sector surplus will equal the global government sector deficit. It is not possible for the global private sector to net save financial assets unless the government sector as a whole runs deficits.

Orthodox policy prescriptions often propose to violate the identity, for example, advocating simultaneous private sector and government sector surpluses despite current account deficits, which is impossible.

Understanding and maintaining the sectoral balances identity in our analyses also helps to maintain stock-flow consistency. Thus, any sector running a deficit must be issuing debt, while any sector running a surplus is accumulating financial claims on the deficit sector(s). This makes it clear that, if the government sector runs a deficit while the private sector runs a surplus, the government must be issuing debt that is accumulated as net financial assets in the private sector.

It also means that we should monitor the private sector's build-up of wealth, unlike orthodoxy which focuses its attention on the government's build-up of debt. In our policy analysis, we must recognise that cutting the fiscal deficit and slowing the government's issue of debt will reduce the private sector's surplus and curtail its accumulation of financial wealth (holding the foreign balance constant). The wisdom of cutting fiscal deficits for some nebulous financial reasons disappears once one takes account of both the currency-issuing powers of governments and the consequences for sectoral balances.

33.3 Government and the Monetary System

A sovereign currency

In all modern, developed, capitalist economies, the central government is the biggest economic entity. Its role cannot be reduced to that of just another interest group, such as a firm or a household. The government has multi-dimensional *sovereign power*, which is not unconstrained, but it does have powers that no other economic entity enjoys.

While orthodoxy (and even some heterodox approaches) imagines that money was invented by private markets in order to reduce the inconvenience of barter-based trade, MMT argues that the macroeconomic analysis of government policy must take account of the difference between a sovereign currency and a non-sovereign currency. The latter applies to members of the Eurozone, for example.

The sovereign government chooses the money of account, issues its own sovereign currency, spends its currency, and enforces most legal contracts in that currency. It also accepts its currency in payment for various obligations, notably taxes, fees, and fines. Thus, there is an important relationship between the government and its monetary system.

Most people would focus on the role played by currency in 'spot' transactions, that is, the use of cash to make purchases. The importance of cash in hand-to-hand transactions varies greatly across modern nations, but has generally declined over time.

The most important role played by currency today is in 'clearing', which usually takes the form of using central bank reserves to clear transactions between banks. The access of the private banks to central bank reserves remains critical to the smooth functioning of the banking system. This is especially true in a crisis. For example, we noted that, during the GFC, the US Federal Reserve Bank had to originate $US29 trillion in loans of reserves to troubled banks and central banks. For this reason, our definition of currency would include cash, treasury coins and treasury and central bank notes, as well as central bank reserves. The definition in different countries might vary by the type of financial assets the government issues.

Fiscal policy

The central bank handles spending for the treasury by crediting private bank reserves. The private banks then credit the accounts of the recipients of government spending. Taxes simply reverse this, with the central bank debiting bank reserves (and crediting the treasury's account at the central bank) and the private banks debiting the accounts of taxpayers.

Thus, government spending and taxing are conceptually separate operations, but most people think that taxes 'pay for' spending. This is not technically true and it creates the misunderstanding that government needs to receive tax revenues before it can spend. Indeed, from inception, taxes cannot be paid in the form of the sovereign currency unless the currency has been issued, usually through past spending (although it could be lent by the sovereign so that taxes could be paid).

In modern nations, taxpayers use private bank deposits to meet those obligations. However, their banks must make payments to the state for their customers using the sovereign currency, usually using central bank reserves. This ensures there is always a demand for central bank reserves. Reserves are created only by central banks; there is no other source of reserves.

This leads to a revised view of government bond sales. Again, from inception, government cannot borrow its own currency in order to spend unless it has already provided its currency to those who are prepared to lend it. Actually, when a government sells bonds, the central bank debits bank reserves by an equivalent amount, whether the banks are the purchasers of bonds, or whether the purchasers are bank customers. Bond purchases (central bank open market purchases or treasury retirement of debt) reverse this operation, with the central bank crediting bank reserves. Rather than seeing bond sales as a way to finance government spending, bond sales and purchases should be considered as part of monetary policy operations.

When government runs a deficit, meaning its spending exceeds tax receipts, there are net credits to bank reserves, and net credits to the deposits of bank customers. In other words, deficit spending by government creates a surplus and thus net financial assets are held in the non-government sector, initially in the form of bank reserves. If the government sells bonds, the reserves are debited so that the net financial assets take the form of government bonds.

According to MMT, fiscal policy can have a powerful impact on the non-government sector because government spending raises private sector income and pushes the private sector's balance toward surplus. Whether the private sector will actually be in surplus also depends on the external balance. In addition, government deficits improve non-government sector balance sheets by providing safe, liquid net financial assets, namely cash, reserves, and bonds.

Not all government spending has the same stimulatory impact. Government spending to directly create jobs, for example, will probably have a greater impact on aggregate demand, than say an equivalent-sized tax cut because those who obtain new jobs will increase spending on private sector output (that creates more jobs through the multiplier).

However, given that the government issues its own currency, it can never run out of finance. It is always possible for government to spend on a sufficient scale to push the economy toward full employment of its labour and capital resources. This does not mean that government necessarily *ought* to do this, since the inflationary effects as well as impacts on the environment depend critically on the composition of government spending, and not just its scale.

The point is that affordability is not the issue, notwithstanding that there are legitimate questions about the impacts of government spending on inflation, financial stability, environmental sustainability, and social stability that must always be considered.

Basing the monetary system on the sovereign currency and imposing liabilities that must be paid in the sovereign currency is the key to creating fiscal and monetary policy space. How much larger the policy space will be depends on the exchange rate regime which is adopted, that is, floating, managed, or fixed.

With a floating exchange rate, there is no risk of insolvency or involuntary default on government debt. Thus, financial affordability is never in question. The real constraint is the full employment of resources, most importantly labour. With a floating exchange rate, the interest rate target can be set to be consistent with domestic policy goals.

With a sovereign, floating, currency, government can remove self-imposed constraints, such as balanced fiscal positions, central bank reserves targets, fixed exchange rates, and perceptions that 'markets' determine interest rates. None of these constraints apply in nations that adopt their own currencies. We discuss different exchange rate regimes in Section 33.6.

Persistent fiscal deficits

While mainstream economists consider the accrual of government deficit and debt to be a financial burden on the population, who will someday have to 'pay these back' via higher taxes, deficits allow other sectors to run surpluses, and the government's accumulated debt is net financial wealth for the other sectors. Further, there is no reason why future generations would need to tax themselves more to 'pay back' the government's debt, which would simply reduce their net financial wealth.

In a normally functioning modern economy, the government runs chronic deficits and the domestic private sector runs surpluses. This is perfectly sustainable – and indeed promotes financial stability in the private sector. The exception would be for a nation that runs chronic current account surpluses, which obviously is not possible for all nations to do simultaneously (since if some nations are running surpluses, there must be others running deficits).

By this we do not mean to imply that a government should never run a surplus. There are two situations in which a surplus might be needed. If a private domestic sector boom is overheating the economy, the government's fiscal position will normally move to surplus, with the private domestic sector running a deficit. This would help to cool the economy. Second if a country has a large current account surplus, which could also overheat the private domestic sector, the government's balance would probably move to surplus.

Central bank creation of reserves to make payments for the treasury frees government from the 'fiscal constraint' that orthodoxy imagines. If government really were constrained by its 'income' (tax revenues) or ability to 'borrow' (sell bonds to savers), it would lose the fiscal power to act countercyclically. In an economic downturn, which leads to falling tax revenue, government would have to behave like households and firms that 'tighten their belts' in recession, adding to the headwinds and worsening the demand gaps. Every recession would become a Great Depression. Thus, freedom from the fiscal constraint is essential to ensure that government does not have to behave like private sector households and firms.

Further, the government's ability to undertake some of its most important functions would be severely limited. Imagine the consequences if the Allied governments during the Second World War had not been able to ramp up spending to fight the fascist Axis Powers (Germany, Italy and Japan). During that war, the US government's spending, for example, grew to 50 per cent of GDP, with its deficit reaching 25 per cent of GDP. This was a common trend for all governments prosecuting the war effort. While a fascist state might be able to conscript soldiers, workers, and industries (as well as enslave large portions of its own and subject populations to work in war production), a democratic nation largely relies on its sovereign monetary power to hire labour and purchase output in pursuit of the public interest. While the advantages are most obvious in times of war, the government's ability to use fiscal power in more normal times is just as essential.

One of the biggest mainstream fears is that a government deficit pushes up interest rates as government borrowing to pay for its spending competes with private demand for savings that are deposited in banks to make loans. This scenario is flawed for two reasons. First, government does not really borrow to pay for deficits, although it might issue bonds when it runs deficits. However, these bond issues are really part of monetary policy and have the effect of removing reserves from banks. They do not drain savings. Indeed, as just discussed, government deficits create net savings and the bonds are accumulated as net financial assets.

Second, the central bank targets the overnight interest rate and can keep that as close to zero as it likes, no matter how large the deficits might reach. The GFC drove this point home as fiscal deficits grew rapidly in most countries, at a time when central banks lowered interest rate targets and rates on sovereign government bonds fell. The orthodox fears are unfounded but persist because mainstream economics does not properly 'account for' government deficits and debt because it does not ensure that its models are stock-flow consistent.

33.4 Monetary Policy

The central bank sets the target overnight interest rate at which banks lend reserves to one another. Its provision of reserves accommodates the demand by banks for reserves, largely for clearing purposes although some countries, such as the US, also have minimum reserve requirements.

The central bank cannot supply fewer reserves than are needed by banks because this would cause them to bid the interbank lending rate above the central bank's target. It can choose to supply more reserves than banks want, but it then must pay the target interest rate on excess reserve holdings or banks would bid the interbank lending rate below the target.

Despite the widespread belief that the central bank is, and should be, independent of the treasury, it cannot in reality be independent. It is the government's bank, making and receiving virtually all payments for the treasury. It must clear treasury cheques (or electronic payments), which means that it must work closely with the treasury to ensure that no payments 'bounce' for lack of funds.

The evidence that this coordination works is that treasury cheques never bounce, nor do taxpayer's cheques due to insufficient reserves being held by their banks (assuming the individuals writing the cheques have sufficient funds in their own bank accounts). Further, the treasury can sell bonds as needed to replenish its balances at the central bank.

Many central banks attempted to control the money supply in the 1970s and early 1980s, but gave up because they were unable to do so, for reasons we discussed earlier in the book. Some mainstream textbooks, however, still claim that the central bank can control the money supply, and, through the imposition of a reserve ratio, the level of credit creation (see below).

Reserves and bond sales

Most economists believe that the treasury sells bonds to finance its spending when its tax revenue is insufficient. As noted, these bonds supposedly compete with private borrowing for a scarce supply of saving and so can push up interest rates. This misunderstands the purpose of bond sales as well as their impact on interest rates.

When government spends, bank reserve holdings at the central bank are credited. When taxes are paid, reserves are debited. Hence, if spending exceeds taxes, reserves are net credited. Normally banks economise on reserve holdings, since earnings on reserves are low with rates being set by the central bank. Banks do not want to hold more than they need for clearing and to meet required reserve ratios (if they exist). Any bank with more reserves than it needs will offer them in the overnight market.

If fiscal deficits are of any significant size, they will create excess reserves in the banking system. This results in the offers of reserves exceeding bids for reserves by banks that need them, which places downward pressure on the overnight interbank lending rate. As modern central banks always target the overnight rate, they will intervene once that reaches the bottom of the target range. To remove excess reserves, central banks sell assets, normally treasury debt. Banks buy treasury bills and bonds using their reserves, which drains the system of any excess reserves.

In practice, the procedures adopted by the central bank, in cooperation with the treasury, are more complicated and more proactive than that. Indeed, in some countries, including the USA, the UK and Australia, the central bank and treasury coordinate activities every morning in anticipation of treasury spending, receipts of taxes, treasury bond sales and redemptions, and other country-specific transactions. The reason is that these activities will affect bank reserves.

The central bank can then plan whether it will sell or buy bonds to offset the impacts on bank reserves of the treasury's actions. Even with careful coordination, it is impossible to perfectly predict all these activities. For example, the treasury does not know precisely how many of the cheques it sent out in payment will be 'cashed', nor exactly how much will be paid in taxes on any given day, so the central bank can adjust its bond transactions during the day.

Many central banks now pay interest on most reserve holdings. This usually eliminates any significant difference for the banks between holding treasury bills or reserves at the central bank. This became essential after the adoption of quantitative easing, which provided banks with high levels of excess reserves that they would not have wanted to hold if they earned zero interest. In such a situation, the overnight rate would be pushed to zero as banks with excess reserves tried to get rid of them. Once central banks began to pay interest, the interbank rate would only fall to the 'support' rate, that is, the rate the central banks pay on reserves.

What does all this mean? From the vantage point of markets, it makes no difference whether bond sales come from the central bank or from the treasury: the functional impact is to offer an interest earning alternative to reserves (that either pay zero, or the support rate paid by the central bank). In other words, the functional impact of bond sales is not to allow the treasury to borrow, but to allow the central bank to implement monetary policy, whose purpose is to hit interest rate targets. It also means that rather than pushing interest rates up, fiscal deficits in the absence of bond sales would normally result in excess reserves and hence push them down.

In summary, there are two quite different methods used to create reserves. First, central banks create reserves either to lend them to banks or to purchase assets. Functionally these two actions are the same. A bank needing reserves can either borrow reserves from the central bank at the discount window (the bank submits an asset as collateral against which the central bank lends reserves), or the bank can sell an asset to the central bank in return for a credit to its reserve account. Note that there is no technical limit to the central bank's ability to create reserves that are desired by banks, since it can always make loans or purchase assets that banks want to sell. The limit is set by the demand by banks for reserves up to the central bank's willingness to supply such reserves, which might include the central bank's evaluation of the quality of assets submitted by banks for collateral or for sale.

Second, the central bank creates reserves to make payments for the treasury. When the treasury issues a cheque for payment, or approves an electronic payment, the central bank credits the reserves of the recipient's bank. The precise procedures adopted for making payments can be varied and complex, but the end result of a treasury payment is always the same; namely, that the recipient's bank account is credited, and that bank's reserves are credited. Note, again, that there is no technical limit to the central bank's ability to create reserves to facilitate treasury payments, and in practice modern central banks always ensure that treasury payments 'clear'.

Central bank creation of reserves through its lending operations ensures that the payments system functions smoothly. When financial institutions face demands to make payments on their own commitments, they can always turn to the central bank to lend reserves so that they can do so.

33.5 Private Banks

Finance

Saving does not finance investment (or consumer deficits, or government deficits); indeed the direction of causation is the reverse. It is investment, plus government deficits and net exports, that generate saving; saving is not the source of finance. Rather, saving is a two-step decision: given income, the saver decides first how much to save, and second, the form in which saving will be accumulated. It is the second decision that plays some role in affecting interest rates. While it is true that at the individual level one can use one's saving to finance spending (such as investment spending), saving is not a source of finance at the aggregate level.

Instead, finance takes the form of credit and typically involves four balance sheet entries. In the case of a loan by a financial institution, the creditor records an increase of assets (the 'loan') and an increase of liabilities (the 'deposit'). The debtor records an increase of assets (the 'deposit') and an increase of liabilities (the 'note' or IOU held by the creditor).

Saving plays no role in this process. Rather, it is a system of debits and credits on the balance sheets of financial institutions. In the real world, the transactions can get much more complicated. (For example, banks today typically bundle many loans together to 'securitise' them, then sell them off to managed money funds.)

Uncertainty and default risk are pervasive phenomena in financial markets. Higher debt ratios in the private sector are associated with greater financial fragility and the possibility of crisis. For that reason, debt matters. The use of retained earnings or sales of equities to finance investment is generally less risky because these sources of finance do not commit firms to future payments.

Likewise, for households, using current income flows to finance spending is less risky than borrowing to finance purchases. As indebtedness grows, the chances that a household will face default and the serious consequences of bankruptcy also increases.

However, the mainstream view, which is still found in most macroeconomics textbooks, is that financial institutions gather the savings deposits of households, firms, governments and foreigners in order to lend those out to investors. Thus, it is claimed that financial institutions perform the function of intermediation between savers and investors. The total supply of finance would then be limited by the thriftiness of savers. This is why there is concern about government borrowing 'crowding out' private borrowing, especially that of firms that would use the finance to invest. The scarce supply of saving meets the demand for finance, which determines the interest rate.

In the orthodox view, it really makes no difference whether investing firms use their retained earnings (their own savings), borrowed funds from financial institutions (intermediated savings) or sales of equity (direct funding from savers). As discussed above, the impact on decision making is the same, given the assumption that no one ever defaults. Consequently, debt does not matter.

Inside wealth versus outside wealth

As noted above, for every non-government sector financial liability issued to another entity within the non-government sector, there is a non-government sector financial asset. These assets and liabilities must net to zero. These are called 'inside' assets and liabilities because they are inside the non-government sector, so inside financial wealth is zero. However, there are also real assets held in the non-government sector which represent net inside wealth. The government sector's liabilities are held by the non-government sector as net financial assets. This is called outside financial wealth. Hence, total non-government sector net wealth is equal to government liabilities (outside financial wealth) plus real wealth.

Credit creation and the money supply

Banks create deposits when they make loans, so the money supply expands endogenously. Reserves do not constrain bank lending, nor for that matter does the amount of vault cash physically held by the bank. In reality loans are just balance sheet entries, as discussed earlier. Banks use reserves for clearing accounts with one another and with the central bank. Since central banks operate with interest rate targets, they must accommodate the demand for reserves for such purposes (or otherwise, the central banks would miss their targets).

Private financial institutions provide the vast majority of those financial assets that we include in our measure of 'money supply' (whether that is narrowly defined to include only cash and demand deposits, or broadly defined to include other short-term liabilities of financial institutions). Only a small percentage of the 'money supply' is issued by government (treasury plus central bank).

However, it must be remembered that banks and other financial institutions do not create net financial wealth, since their assets are the debts of the non-financial sectors and their liabilities are the assets of the non-financial sectors. Only the government can create net financial wealth for the non-government sector.

In the orthodox textbook view, which still prevails, banks take in deposits, hold some in reserve, and then lend the excess. At the aggregate level, there is a deposit multiplier in a fractional reserve system. Since banks need reserves to make loans, it is the central bank that ultimately determines how many new deposits can be created. In other words, the central bank controls the growth of the money supply through its control over reserves. Most orthodox economists now realise that central banks target interest rates, and they do not (perhaps cannot) control the money supply. Yet, many of the orthodox textbooks still teach the deposit multiplier story.

33.6 Trade and Exchange Rates

The external balance is the spending by the rest of the world on goods and services produced in the nation, as well as the flow of factor payments received from the rest of the world, less national spending on goods and services from abroad plus factor payments to the rest of the world. This is summarised in the current account balance, which can be negative or positive:

Current account balance =

(Exports of goods and services + Factor payments received from rest of world)

− (Imports + Factor payments paid to rest of world)

A current account deficit means that the rest of the world is running a surplus (its spending is less than its income), while a current account surplus means that the rest of the world is running a deficit. There is a stock counterpart: a current account surplus means that the nation is accumulating financial claims on the rest of the world; a current account deficit means that the nation is accumulating indebtedness to the rest of the world:

Current account surplus → Accumulate claims on rest of world

Current account deficit → Increase indebtedness to rest of world

To recap, at the aggregate level:

Domestic private balance + Domestic government balance + Foreign balance = 0

The only way that a nation can simultaneously run a private sector surplus and a balanced government fiscal position, is by running a current account surplus (so that the foreign balance is negative as the rest of the world runs deficits against the nation).

Exchange rate regime

Finally, we return to the exchange rate regime adopted by the sovereign currency issuer. If the domestic currency is fixed to a precious metal or a foreign currency, the domestic policy space is reduced. The nation needs to operate its economy in such a way as to ensure a positive flow of the metal or foreign currency. This can be done by achieving positive net exports (or more broadly through a positive current account balance). That usually means that it needs to suppress imports and encourage exports. Both strategies are facilitated through the adoption of domestic austerity policies, thereby keeping income down, which reduces consumption, including spending on foreign output (imports), and also cheapens domestic production costs (encouraging exports). It is usually believed that a positive trade balance keeps a currency strong, which prevents downward pressure. The country might also keep interest rates high for a similar reason.

If the nation cannot run a positive external balance, it can still obtain foreign currency through borrowing, by perhaps issuing debt in a foreign currency. However, with a fixed exchange rate, default on government's promises becomes a danger. First, it might have to devalue the currency if it fears it cannot cover the demand for foreign exchange at the fixed rate. Second, if it has issued debt in a foreign currency, it might have to default on the promise to deliver the foreign currency in payment.

As fears grow that a government might default, there can be a run on the domestic currency, which takes the form of submitting the domestic currency for redemption in foreign currency (or precious metal). Just as no bank can survive a bank run unless it has sufficient reserves to cover its promises to convert, no country that has pegged its currency can survive a run without sufficient reserves. The run can only be stopped if the supplier of reserves acts as a lender of last resort. Normally, countries that peg cannot rely on the issuer of the reserve currency to come to their rescue.

Pegging a currency essentially reduces a country to the status of a user of the foreign currency. Even if the government continues to issue its own currency, it needs the foreign currency (or gold) to protect it from default.

The sovereign government might instead choose a managed exchange rate. In this case, the government chooses to maintain the exchange rate against a foreign currency (or precious metal) within some range. It tries to build the expectation that it will be able to keep the exchange rate stable. However, it might voluntarily choose to change the exchange rate, or it might be forced to do so. Since it has not promised a fixed exchange rate, this is not a default. (Although if the government issues liabilities denominated in foreign currency, it can be forced to default.) However, the danger of building the expectation of stability is that if holders of liabilities denominated in the currency begin to doubt that the exchange rate will be stable, they will put pressure on it to move. This expectation can thus become self-fulfilling, so that the desired stability cannot be achieved.

Floating the domestic currency eliminates the possibility of a default on commitments (so long as the government does not issue debt denominated in foreign currency). It also makes it less profitable to speculate against a currency. In the case of fixed or even managed exchange rates, speculators can take 'short' positions – bets that the government cannot obtain sufficient foreign reserves to keep the exchange rate stable. (The speculator will sell the domestic currency to buy the reserve currency which puts downward pressure on the exchange rate.) If the shorts become big enough relative to the government's reserve holdings, these speculative bets become virtually a sure thing, since the government is forced to use its scarce foreign reserves to purchase its own currency to address the downward pressure on the exchange rate. The speculator's gain is the government's loss.

If on the other hand, the government floats the currency, it does not have to intervene. Knowing this, it does not (necessarily) operate its economy in a manner that would allow it to accumulate foreign reserves.

Free from financial constraints, the sovereign government can use the domestic policy space to pursue the most important public purposes: full employment; provision of adequate food, clothing and shelter for all of its population; universal access to healthcare, education, and culture; aged care, maternal care and childcare; protection of the environment; public infrastructure investment; consumer protection and regulation of private enterprise; promotion of civil rights and equal opportunity; pursuit of domestic safety and protection of legal rights; and provision of assistance to other nations. No government with its own sovereign currency can legitimately claim that any of these activities 'costs too much'. These are financially affordable. The only question is whether there are sufficient resources that can be mobilised to pursue these goals.

This does not mean that every nation should float its exchange rate; however, all sovereign nations should carefully evaluate the constraints that arise from pegging or managing exchange rates.

Conclusion

Over the past half century mainstream macroeconomics has become increasingly devoid of relevance to our understanding of how modern monetary economies operate. In part this is due an increased emphasis on mathematical models underpinned by simplistic assumptions that reduce human behaviour to that of a 'rational agent' who maximises simple goals in at worst a 'risky' economic environment.

How modern economic institutions such as the central bank and treasury actually operate is either misrepresented or largely ignored. In addition, in an attempt to ground macroeconomics on what are believed to be good 'micro foundations', modern mainstream theory uses constructs such as the single representative agent to stand in for society. This has led to a set of claims which are blighted by the 'fallacy of composition'.

Taking all of these features together, modern macroeconomic theory has blinded its proponents to such a degree that they could not anticipate the Global Financial Crisis, and instead, led them to propose policies that hastened the onset of the crisis. Even after the crash, they supported policies that have delayed recovery and increased the chances of another crisis.

The MMT framework developed in this textbook is based on a fundamentally different approach to macroeconomics. As has been demonstrated, MMT practitioners did see the crisis coming. Those who foresaw the crisis adopted a more realistic approach to human behaviour, which built on the work of giants such as Marx, Veblen, Keynes and Kalecki. Humans are social animals whose behaviour is shaped by institutions, by classes, by custom and by social standing. Their decisions are made in conditions of true uncertainty, where rational calculation of the probabilities of different outcomes is not possible.

Further, those economies that are the subject of the analysis of macroeconomics are *capitalist* economies organised around what can be called 'monetary production'. The goal of private production in these economies is to 'make money', that is, to accrue profits in the form of money claims. Money cannot be 'neutral' in these economies, for the production process itself is oriented toward money profits.

As the three great economists Marx, Veblen and Keynes would argue, capitalist production begins with money with a view to ending with more money. The prospects for profitable production are always uncertain. If production were merely a technical matter, there would be little uncertainty: combine the 'factors of production', following well-known templates, to produce 'widgets' to be sold at market clearing real (relative) prices.

However, capitalist production is much more complex. The capitalist must first obtain money to hire the factors of production, which will probably entail entering into debt commitments (loan agreements) that require repayment with interest. Whether these commitments can be met will depend on the revenues to be generated in the future. This is particularly problematic when it comes to the investment decision, in which expensive, long-lived capital assets are purchased with debt that might have to be serviced for five, ten, or even twenty years or more. The investing firm has no crystal ball that will reveal its income flow over such long periods of time.

For this reason, debt matters. Holding highly liquid assets as well as building in margins of safety is prudent behaviour in such conditions. The preference for liquid positions results from uncertainty over future outcomes, and the degree of liquidity preference is subject to waves of optimism and pessimism over the course of a business cycle. A boom can generate euphoria and willingness to discount downside risks on both the part of the borrower as well as the lender. Financial fragility can grow, making a crisis more likely, even as borrowers and lenders embrace the view that the possibility of crisis is becoming increasingly remote. Policymakers promoted this false feeling of security by embracing the 'Great Moderation' prior to the GFC and threaten to repeat this by disingenuous rhetoric about the 'new economy'.

Readers of this textbook who have engaged with and critically evaluated the material are less likely to be duped. They will recognise that market forces need not be stabilising. Indeed, as Minsky argued, market forces can be destabilising and even explosive from time to time. Instead, it is the institutional structure imposed on markets that can constrain the instability. However, market participants react to apparent stability in ways that reduce the effectiveness of the constraints. Economists and policymakers need to be sceptical of claims of the rise of a 'new era' of stability.

Further Reading

Mitchell, W.F. and Fazi, T. (2017) *Reclaiming the State: A Progressive Vision of Sovereignty for a Post-Neoliberal World*, London: Pluto Books.

Visit the companion website at www.macmillanihe.com/mitchell-macro for additional resources including author videos, an instructor's manual, worked examples, tutorial questions, additional references, the data sets used in constructing various graphs in the text, and more.

INDEX

J

K

L

M

N