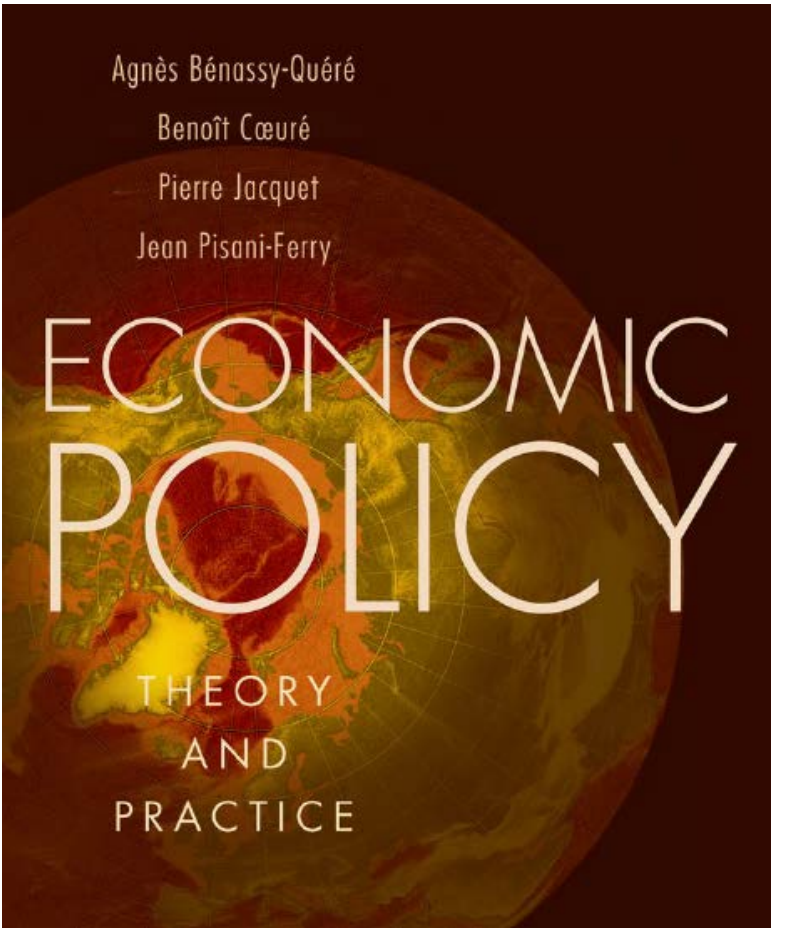
Economic Policy

Theory and Practice

Agnès Bénassy-Quéré, Benoît Cœuré,
Pierre Jacquet, and Jean Pisani-Ferry



OXFORD
UNIVERSITY PRESS

2010

Foreword

This book is a book I would have loved to write. Indeed this is a book I long wanted to write. I wanted to do so out of guilt. For a long time, I have felt that my graduate textbook written with Stan Fischer sent the wrong message. We had made the choice to present models and their logic, rather than their applications. The justification was a perfectly good one, namely that we wanted to show the intellectual structure of macroeconomic theory first. But, de facto, the lack of serious empirics sent another message: That theory was largely divorced from practice, and from facts. That message is wrong: Theory without facts is much too easy, and of very little use. I also wanted to write such a book out of a desire to share with students my excitement about moving between theory, facts, and policy. It is traditional to do so in undergraduate textbooks, at least in the United States. Those textbooks are full of discussions about policy debates—about the effects of policy choices on the economy. I thought it would be even more fun to do so with graduate students, who have more tools, both theoretical and econometric, at their disposal.

Agnès Bénassy-Quéré, Benoît Cœuré, Pierre Jacquet, and Jean Pisani-Ferry have beaten me to it. I am happy they did so, because they have done a better job than I could have hoped to.

To give a sense of what they have achieved, I shall take one example, the creation or reform of fiscal frameworks like the European Stability and Growth Pact (SGP). To come to an intelligent set of recommendations, think of all the elements you need to put together:

- You need to understand what sustainability means in theory and in practice, what the costs of not abiding by sustainability are, and how to

assess sustainability. When does a debt-to-GDP ratio become truly excessive? What happens then? How fast can you reach that threshold? How fast can you move away from it?

- You need to understand the long-term effects of deficits and debt on output and its composition. How do deficits and debt affect output in the medium and the long run? How do they affect the interest rate, the net foreign debt position, the capital stock? What is the cost in terms of lost consumption in the future? Which generations gain, which generations lose?
- You need to understand the short-term effects of deficits, and how counter-cyclical fiscal policy can help in the short run. Do deficits affect activity in the same way, whether they come from tax cuts or spending increases? How important are expectation effects? Can the anticipation of large deficits in the future lead to a decrease in consumption and investment, and a decrease in output today? When is this more likely to happen?
- You need to understand the macroeconomic costs of decreased policy flexibility. Are constraints on deficits and debt consistent with an appropriate response of fiscal policy to shocks? What explains sustained divergences within the euro area during the first 10 years? Were such divergences avoidable? Then you need to determine whether and to what extent fiscal policy is the right tool to deal with country-specific shocks, and to what extent it can (should) substitute for the lack of an independent monetary policy. Finally, you need to figure out how much policy space is left to governments after they have fought the new great recession and rescued their banks.
- You need to think about how to define the rules in practice. How should debt be defined? How should implicit liabilities, coming from social security and other promises to future generations, be treated? If rules are defined in terms of deficits and debt, what are the most appropriate definitions of the two concepts for the question at hand? How should rules deal with privatization revenues? Should rules apply to gross debt or to net debt? Should the budget be separated between a current account and a capital account? Should the deficit rules apply only to the government current account? Should rules be enforced by politicians or by independent committees?
- You need to think about political-economy issues. Why are rules needed in the first place—to protect people from their governments, or to protect the governments from themselves? How can a particular set of rules be manipulated or distorted by a national government? How will sanctions against a misbehaving government be imposed? Will these sanctions be credible *ex ante*? Is international coordination, such as in the G20 framework, an advantage or a diversion from every government's duties?

To answer these questions, you need many conceptual tools. Among them are: A dynamic general equilibrium model with overlapping generations; a model of short-run fluctuations with careful treatment of expectations; political-economy models to think about the case for rules; agency models to help you think about the design of specific rules. In each case, with guidance from theory, you need to look at the evidence, so as to get a sense of which theoretical arguments are more relevant. This is not easy to do. Courses will typically give you the theoretical tools, without much incentive to apply them, and leave you to use them on your own, without much practical training. This is not what this book does. It motivates you to use tools, gives you the tools, and shows you how they can be employed.

Last but not least, this book is among the very first that offer students a rigorous and comprehensive treatment of the financial crisis and the great recession that followed. The authors do not try to cast a veil over the conceptual difficulties economists face when they reflect on the causes of the crisis, on the limitations of traditional approaches that the crisis has uncovered, and maybe the excessive faith in theory, and on the need for more theoretical work to understand better the crisis and make sure it does not happen again. But they do not throw the baby out with the bath water and claim that economists have "mistaken beauty for truth," as was suggested by Paul Krugman. On the contrary, they show how existing theories can be used, cross-fertilized, and placed in a historical context to understand the crisis better. This is the way forward.

In short, this book trains you to be a good macroeconomist—a good economist. It instills the right attitude, and gives you the right methodology: To build solidly on theory, to use the theory to look at the data, and then to go back and forth between the two until a coherent picture forms. As I was reading the book, I felt again the excitement that comes with doing research on macroeconomics. I hope this excitement is contagious, and I wish you a very good read.

Olivier Blanchard
Massachussetts Institute of Technology and
International Monetary Fund

Preface

This is a book for all those interested in what shapes, or should shape, economic policy: The major stylized facts that capture the messages from history; the theories that help us understand these facts and represent the impact of policy decisions; the controversies surrounding policy choices; the institutions that contribute to determining them; and, last but not least, how experience, theories, and institutions interact.

We have been teaching the material that forms the basis of this book in a graduate seminar at the École Polytechnique in Paris since 1998, and also at Sciences Po, École Normale Supérieure, École des Ponts–ParisTech, and Université Paris-Dauphine. In 2004 a first book in French arose from this experience, followed by a second edition in 2009. This English edition presents completely revised material, drawing on our experience with previous editions and on the policy lessons learnt in the recent global crisis.

Preconditions for using this book are limited, because we start from facts, introduce theories as needed, and keep formulas in boxes. Practitioners and observers will find what they need to understand actual policy issues and discussions. However, graduate students more familiar with models will also learn how to link leading-edge research to concrete policy developments.

This book also aims to eschew cultural bias. Our analysis starts from policy questions in Europe, the US, and the emerging world, and our examples are taken from around the world.

Why This Book Is Different

This book is based on the premise that the disconnect between theory and practice is detrimental to both good policy and good research. It posits that going back and forth between practice and theory enlightens practice and helps construct better theories.

We have been vindicated in this belief by what we have learnt from personal experience. Each of us has engaged at times in academic research, policy advice, and policymaking, at a national, European, or international level. This has changed the way we understand and use economic theory.

This is why we have embarked on this project, and aimed to provide a systematic and theory-driven approach to economic policymaking. Economic textbooks typically cover economic theory in a given field—macroeconomics, microeconomics, finance, international trade, etc. Real-life stories are often recounted to illustrate theoretical results. However, the representation of economic-policy instruments and of the decision-making process remains very rudimentary and abstract. Conversely, there are many excellent essays on economic policy, but they are more concerned with describing the ebb and flow of new ideas and institutions than with discussing their theoretical underpinnings. Our book aims to fill that gap.

The result is admittedly an unusual book. The blend of facts, theory, and practice departs from what is found on most courses. We regard this structured eclecticism as the book's comparative advantage. Many of our students and colleagues have commented that what they have read in the book could not be found elsewhere.

Our aim has been first and foremost to help readers build bridges between the elegant theoretical constructs taught in universities or discussed in seminars and the mere plumbing that constitutes the daily life of economic policymaking in ministries, central banks, and international organizations. Usually, economists begin by learning the former and discover the latter only later in their career. We trust that this book will make a significant contribution to preparing students for the challenges of effective economic policymaking, and will increase the policy value of their academic background.

How to Use This Book

This book summarizes the main theoretical and empirical instruments, old and new, that are relevant to addressing real-life policy issues; it explains how these instruments can be used to identify policy trade-offs and guide policymakers' choices; and it discusses the theoretical uncertainties, blind spots, and controversies that warrant humility and caution when formulating policy advice—and that make the job of economists so challenging and rewarding.

There are eight chapters. The first two chapters set out the general framework of economic policymaking. Chapter 1 describes the methodological

foundations and details of the toolbox which will be used in the rest of the book. Chapter 2 adds a note of caution: It outlines the limits of government intervention in the economy and the political-economy arguments which may render it sub-optimal. Chapters 3 to 7 cover in turn five domains of economic policy: Fiscal policy (chapter 3), monetary policy and financial stability (chapter 4), international capital movements, the choice of exchange-rate regimes and exchange-rate policy (chapter 5), long-term growth policies (chapter 6), and tax policy (chapter 7). Finally, chapter 8 covers the 2007–09 global crisis and its lessons for economic policy.

Each of the five central chapters (3 to 7) is structured in a similar way: Stylized facts are taken from recent economic history, then the theoretical tools available to policymakers and which they should be mastering are explained, and finally the main policy options are presented. There are many cross-references between the five chapters, but they are written in such a way that each of them can be read on its own.

Economists are often blamed for resorting to technical vocabulary as a way of protecting themselves from inconvenient questions. We have tried to unveil the—often simple—concepts behind complicated or abstract expressions such as the output gap, welfare losses, or rational expectations. A detailed index lists all these concepts, and points to the place in the book where they are defined, explained, and illustrated. Additionally, there are extensive bibliographical references so that the reader can dig further into any of the issues covered.

This book is by no means comprehensive. Individual behavior, constraints, and incentives are deliberately introduced only insofar as they help understanding of macroeconomic issues. We have thus chosen not to address a number of otherwise important areas of economic policy, such as competition policy, procurement rules and auction schemes, public or private ownership of companies, health care and pension planning, and what has generally been referred to as "mechanism design" by Nobel Prize winners Leonid Hurwicz, Eric Maskin, and Roger Myerson—that is, designing efficient solutions to collective-decision problems. We have also decided not to write specific chapters on international economic policy, international trade, regional (and especially European) integration, or the management of local governments. Chapter 2 summarizes what economic theory has to say on the assignment of policy instruments to different levels of government, and on the difficulties of global governance. However, in any policy domain, some levers are global, some are regional, some are national, and some are local, and we have therefore addressed them in conjunction in each of the five central chapters.

What Has Changed with the Crisis?

As a science, economics has always leapt forward when new facts could not be explained by the prevailing theories, or when economists had to understand

why their advice had failed. Keynesianism triumphed in the aftermath of the Great Depression, which helped in understanding that aggregate demand mattered; the so-called "rational-expectation revolution" of the 1970s prospered when it appeared that Keynesianism could not eliminate stagflation. This is why economics has been striving and will continue to strive to be an intellectual discipline. We have done our best to recognize this and incorporate in each chapter the latest theoretical developments.

However, the global economic, financial, and social crisis of the late 2000s has raised disturbing questions. It has forced governments and central banks to take bold, unprecedented measures, and to radically revisit their policy frameworks. It initially left the economics profession remarkably silent, as if mesmerized. Its impact on economic thinking may one day be compared with that of the Great Depression—or it may not: Only history will tell.

While waiting for that judgment, two kinds of lessons should be drawn. First, operational features of the economy which were considered mere technicalities prior to the crisis, such as liquidity provision or capital requirements for banks, have proven critical to the continuation of economic activity and should therefore be part of mainstream economic knowledge. Second, basic theoretical features, such as moral hazard, market efficiency, or the assignment of monetary policy instruments, have proven more elusive than previously thought, and deserve fresh discussion in the light of this crisis. This has been included in this book where relevant, in particular in chapter 4, which deals with monetary policy, and in chapter 8, which specifically addresses the causes and consequences of the crisis.

Conclusion

We express our gratitude to those who have encouraged us and who have helped make this adventure a reality. We owe a lot to our students, whose questions and criticisms have greatly improved the relevance, accuracy, and legibility of this book. We also thank our colleagues and friends who have commented on previous versions specially Laurence Boone, Benjamin Carton, Elie Cohen, Anne Epaulard, Martin Kessler, Jean-Pierre Landau, André Sapir, Paul Seabright, Nicolas Véron, Charles Wyplosz and our development editor Bill Amis. We reiterate our thanks to Olivier Blanchard, whose work has often inspired us, for having agreed to write the foreword to this book. Last but not least, we thank our families for their patience and support for this seemingly never-ending (and probably ongoing) endeavor.

Agnès Bénassy-Quéré, Benoît Cœuré, Pierre Jacquet,
and Jean Pisani-Ferry

Contents

1	Concepts	3
2	Economic Policy in a Complex World	62
3	Fiscal Policy	152
4	Monetary Policy	238
5	International Financial Integration and Foreign-Exchange Policy	339
6	Growth Policies	436
7	Tax Policy	537
8	Economic Policy and the 2007–09 Crisis	617

1

Concepts

1.1 A primer on economic policy
1.1.1 The economist and the Prince: Three alternative approaches
1.1.2 What do policymakers do?
1.2 The whys and hows of public intervention
1.2.1 The three functions of economic policy
1.2.2 Why intervene?
1.3 Economic policy evaluation
1.3.1 Decision criteria
1.3.2 *Ex post* evaluation and experiments
1.3.3 Collateral effects
Conclusion
References

> Practical men, who believe themselves to be quite exempt from any intellectual influences, are usually the slaves of some defunct economist. Madmen in authority, who hear voices in the air, are distilling their frenzy from some academic scribbler of a few years back. I am sure that the power of vested interests is vastly exaggerated compared with the gradual encroachment of ideas. Not, indeed, immediately, but after a certain interval; for in the field of economic and political philosophy there are not many who are influenced by new theories after they are twenty-five or thirty years of age, so that the ideas which civil servants and politicians and even agitators apply to current events are not likely to be the newest. But, soon or late, it is ideas, not vested interests, which are dangerous for good or evil.
>
> John Maynard Keynes (1936), chapter 24, paragraph 5

The last sentences of *The General Theory of Employment, Interest and Price* by the famous British economist are a fetish quotation for economists, who take them as an acknowledgment of their social role. Yet they also express

the complexity of the links between theory and economic policy. They suggest that economic expertise cannot be regarded as the servant of political decision. Rather, it influences it, although in an indirect way and with delay.[1]

However, Keynes also expressed detached irony about the economists' pretence to determine the policymakers' choice:

> If economists could manage to get themselves thought of as humble, competent people on a level with dentists, that would be splendid.
>
> John Maynard Keynes (1931)

The interaction between economic ideas and political motivations was aptly characterized in the classics as *political economy*.[2] This type of interaction between power and knowledge is certainly not specific to the economic discipline. It arises in all fields where public decision relies at least partially on scientific or technical expertise. For reasons we develop later in this chapter and throughout the book, however, it is more pronounced in economics and more general in the social sciences than, say, in geology or biology.

This chapter provides both an introduction to, and a first discussion of, the main themes of economic policy analysis. It does not enter into the analysis of the specific policy domains and issues that are the topics of the following chapters, except by way of illustration. We start in section 1.1 with a discussion of the various approaches to economic policy an economist can adopt. In section 1.2, we discuss the arguments for and against public intervention, both from a micro- and a macroeconomic standpoint. Finally, section 1.3 is devoted to the evaluation of economic policy choices and deals both with criteria and instruments.

1.1 A Primer on Economic Policy

1.1.1 The economist and the Prince: Three alternative approaches

The economist can adopt diverse attitudes vis-à-vis political decision: she or he can limit herself to studying the effects of public choices on the economy (*positive economics**); she can seek to influence them through making

1. Keynes himself did not escape this rule: his key recommendations were implemented only after World War II.

2. The meaning of this expression has changed over time. In an historical sense (for example the one Jean-Jacques Rousseau (1755) uses in the "political economy" entry of the *Encyclopaedia* of Diderot and d'Alembert), "political economy" was equivalent to "general economics" as opposed to "home economics." Jean-Baptiste Say, Adam Smith, David Ricardo, and Karl Marx, among others, used the expression in this way. It kept this meaning in some countries until the end of the twentieth century. In English, however, *political economy* has been replaced by *economics*. In the US, and later in Europe, a different meaning started to emerge in the 1960s as a strand of research began to explore systematically the political determinants of policy decisions. The corresponding approach was first called *new political economy* but became known as *political economy or political economics*. We follow this usage.

recommendations that draw on her expertise (*normative economics**); or, finally, she can take political decisions as a topic for research and endeavor to identify and explore the determinants of economic policy decisions (*political economy**).

All three approaches coexist in today's economics.

a) Positive economics

In positive economics, the economist takes the point of view of an outside observer and aims at determining the channels through which public decisions affect private behavior. For example, she analyzes the effects of a tightening of monetary policy, an increase in public expenditure, a tax reform, or a new labor market regulation. Economic policy choices are regarded as entirely *exogenous** meaning that they impact on economic variables such as prices, output, or employment without being influenced by these variables.

Positive economics therefore approaches economic policy with the same concepts and the same methods as those used to study other economic phenomena: There is hardly a difference between studying the effects on nonfinancial agents of a rise in the rate at which the central bank lends money and analyzing the effects of an exogenous rise in the risk premium required by banks for lending to private agents; similarly, the effects of a rise in the minimum wage can be analyzed within the same framework and with the same tools as those of a strengthening of the bargaining power of trade unions.

b) Normative economics

The second approach is called normative economics. The economist here adopts the posture of an adviser to a supposedly benevolent Prince—or to any other political master—and examines which set of decisions can best serve explicit public policy purposes, such as reducing unemployment, improving the standard of living, or safeguarding the environment. The public decision-maker is regarded as a social planner, and the economist as an engineer who tells him or her how to select adequate means for reaching certain ends. Economists are certainly not short of policy advice and they generally do not need a request from the Prince to express their views. However, even in this case they make explicit or implicit assumptions about social preferences that cannot be derived solely from economic theory.

Normative economics relies on the knowledge base of positive economics in order to assess the effects of different possible decisions. However, it also requires other instruments, because deciding on a recommendation requires a metric within which to compare alternative situations. Assume that a government wants to reduce unemployment, and suppose that two competing policies may lead to this result, but at the price, for the first one, of a lowering of the employees' average wage income and, for the second one, of

an increase in wage inequality. Choosing between these two solutions requires assessing the social cost of each of them against the social benefit of lowering unemployment. This implies defining a preference order between situations each characterized by the unemployment rate, the average wage income level, and a measure of inequality. Constructing such a ranking raises considerable conceptual and practical difficulties.

Furthermore, normative economics frequently implies giving up the *first-best** solution that would be reached in the absence of informational, institutional, or political constraints for a *second-best** one that respects those constraints.[3] Let us take the example of CD burners, which allow each consumer to copy his or her preferred music. It is reasonable to assume that this technological innovation improves social welfare, but it benefits consumers at the expense of artists, whose CD sales have fallen.[4] Its benefits are thus unevenly distributed. A "Big Brother" who might closely observe the behavior of every music amateur could, for example, learn that the consumers who copy and swap music are mostly urban rock fans between 15 and 35 years old. The first-best policy would thus be to levy a lump-sum tax on them and to compensate the rock musicians for the loss they have suffered, without affecting the rural population or classical music amateurs or those over 35. The social benefit of the innovation would remain but its distribution would be corrected.

However, this solution is out of reach, both in view of the information it would require and because of the legal obstacles it would raise. In practice, a possibility is to tax the sale of CD burners, with the consequence that the benefits of the innovation will be reduced. Another second-best solution, which has been applied in countries like Belgium and France, is to tax the sale of blank CDs and to transfer the product of this levy to the relevant Music Publishers Association; however, this also involves several disadvantages: consumers who buy CDs to store their holiday pictures or to duplicate their music CDs and listen to them in their car are taxed without motive; and the Music Publishers Association may redistribute the product of the levy to all authors, including those whose music is not copied. A number of new distortions are thus introduced in the name of correcting a distortion. That this improves welfare cannot be taken for granted.

Economists involved in public decisions usually face many such constraints. The question they face is not "how can unemployment be reduced?," but "in view of the stance and prejudices of the main players—government departments, majority and opposition in Parliament, and various stakeholders—what is the most cost-effective proposal consistent with the government's overall policy philosophy and commitments already publicly undertaken?." This second question obviously is a very weak version of the first

3. This terminology is taken from welfare economics, which is introduced in section 1.2.2 of this chapter.

4. We neglect music publishing companies here in order to focus only on artists and consumers.

one; but major economic decisions are very often taken this way. Economists may understandably be tempted to abstain from participating in such decisions, but as Herbert Stein, a chairman of the Council of Economic Advisors under US presidents Richard Nixon and Gerald Ford, used to say, "Economists do not know very much [about economics. But] other people, including the politicians who make economic policy, know even less" (Stein, 1986, p. xi). Returning to the ivory tower may thus be an undesirable option.

Second-best recommendations, nevertheless, raise important difficulties. The second-best optimum can in fact be inferior to the initial situation in welfare terms. A standard example can be found in trade policy: liberalization on a regional basis can divert trade from an efficient global producer to a less efficient regional partner, which worsens the allocation of resources in comparison to a situation of uniform tariff protection.[5] What is perceived as a small step in the right direction therefore does not necessarily improve upon the status quo, on the contrary, it can reduce welfare. Following Kemp and Wan (1976), many studies have, however, found that preferential trading arrangements can in fact be welfare improving and contribute to multilateral liberalization. So neither the blind pursuit of regional trade liberalization nor its outright rejection are justifiable attitudes.

Beyond this disturbing result, modern public economics emphasizes the equally formidable difficulty associated with the existence of *asymmetric information** between the public decision-maker, the agents in charge of implementing policies, and those who bear the consequences. Not unlike Soviet central planning, the traditional approach of economic policy postulated that the decision-maker had perfect information (in fact, he or she was frequently assumed to know better than private agents) and perfect control over the implementation of his decisions. The reality, of course, is that the decision-maker has both an incomplete knowledge of reality and an imperfect command of policy implementation. Take the regulator in charge of a specific sector, say telecommunications. He gets information on costs, returns on investment, or demand elasticity largely from the operators whom he is responsible for controlling. For the latter, this information has strategic value. They have every reason not to be fully transparent or to provide biased information. When dealing with them, the regulator therefore suffers an informational disadvantage, even when he supplements the information provided by the regulated companies with indirect indications derived from observing market prices and quantities.

Likewise, government bodies responsible for policy implementation commonly fail to communicate adequately either regarding information from below or instructions from above. For example, even if local civil servants from the labor ministry have detailed knowledge of the employment situation in their area, the minister in charge may not have accurate overall information, which obviously affects the quality of his or her decisions. Reciprocally, the

5. This classical result of trade policy theory was first established by Jacob Viner (1950).

minister's policy may not be completely known and understood by all the civil servants in his or her ministry, and this affects its implementation and effectiveness.

The importance of information asymmetries for private markets was first highlighted in research by 2001 Nobel laureates George Akerlof, Michael Spence, and Joseph Stiglitz, but it was Jean-Jacques Laffont[6] who first pointed out their consequences for public economics. This led him to initiate research on the design of contracts that encourage agents to reveal the information they have rather than keep it for themselves (thereby inducing regulators to take inappropriate decisions).

In Europe, the allocation of third-generation (3G) mobile telephone licenses in 2000 provided a vivid illustration of those difficulties. While licenses were granted at no cost in Asia, most European governments decided to sell them. Setting a price was particularly difficult, however, in the absence of accurate information on fixed costs, variable costs, and future demand. Moreover, telecoms companies had every incentive to overestimate costs and underestimate revenues. Some countries, such as Germany and the UK, chose to allocate licenses by competitive bidding. Since candidate operators set their offers according to their own cost estimates, their bids were expected to reveal the information they had. Actually, the operators grossly overestimated future revenues from 3G telephony and underestimated its development costs, but the bids nevertheless revealed the information they had at the time of bidding.[7]

For the three main reasons given here—the need to define policy objectives and to trade-off for alternative objectives, uncertainty about the correct decision in a second-best world, and information asymmetries—normative economics is fraught with difficulties that positive economics does not need to address.

c) Political economics

The third approach is what is called today *political economics** or political economy. Like positive economics, of which it can be regarded as an extension, the political economy approach refrains from making prescriptions and takes the viewpoint of an external observer. However, instead of considering the political decision-makers' behavior as exogenous, it treats it in the same way it treats private agents' behavior, i.e., as *endogenous** (determined by the state of the economy itself). The government is therefore no longer regarded as a *Deus ex machina* that monitors and steers the private economy in the name of the general interest but, instead, as a machine directed by politicians, i.e., by

6. Jean-Jacques Laffont, a French economist who died prematurely in 2004, initiated the integration of asymmetric information into public economics and applied it to the design of efficient regulation policies.

7. For further information on the bidding mechanisms and on the allotment of the 3G licenses, see Klemperer (2004).

rational players whose behavior follows specific objectives and faces specific constraints. The simplest models of politically motivated behavior often draw on the simplistic assumption that the politicians' only objective is to hang on to power, and therefore to maximize their reelection chances. However, more elaborate models also take into account the need to fulfill electoral campaign pledges (which become a constraint after election) and partisan preferences, which may dwell on the need to maintain long-term relationships within a social group or—at the extreme—on corruption and bribery. The political economy approach also endeavors to represent the behavior of technocrats within government or of those in charge of public agencies (central banks, independent authorities, international institutions), and to determine how the governance and the mandate of these institutions influence economic performance.

Political economy does not exclude normative judgments, but it does have implications as regards their scope. James Buchanan, one of the initiators of modern political economics, claims that such judgments are valid only if applied to the framework (often called *policy regime**) which determines economic policy: the constitution, and more largely all the rules, procedures, and institutions surrounding economic policy decisions. To draw on a distinction introduced by Robert Lucas, the choice of an economic policy regime involves normative considerations, but the actual economic policy *decisions* are the result of political processes within the framework of this regime. It would therefore be pointless to exercise normative judgment on what must be regarded as endogenous variables. According to Buchanan, "the object of economic research is 'the economy,' which is, by definition, a *social organization*, an interaction among separate choosing entities. [. . .] there exists no one person, no single chooser, who maximizes *for* the economy, *for* the polity [. . .] That which emerges [from the decision-making process] is that which emerges from results, and that is that" (Buchanan, 1975, pp. 225–26.). The role of the economist is to study the functioning of these processes and the incentives they create for public decision-makers. It is to discuss whether these incentives create a political bias or help align the outcome of the decision process with the public interest. It is not to give advice to the Prince or his marquises.

During the last decades of the twentieth century, the political economy approach was strengthened by two concomitant developments. First, the theory of *rational expectations**[8] developed in the 1970s (in particular by Robert Lucas) emphasized that private agents do not react to stimuli as automatons, but rather use their reason to anticipate policy decisions. A good example of such behavior is provided by exchange-rate crises.

8. Expectations are said to be rational when economic agents exploit all available information on the functioning of the economy and the variables relevant for their decisions and form the best possible forecasts. In the framework of a model, the rational expectation of a variable is the forecast that can be made within the model by using all available information on exogenous variables.

As developed in chapter 5, such crises can only be understood by taking into account the strategic game between private speculators and official authorities. These crises often occur because private agents know the public decision-makers' preferences, or at least guess what they are, and therefore can assess the probability of a currency devaluation. While not directly related to the political-economy approach, the theory of rational expectations thus challenged the idea that the State dominates and steers the private economy. It resulted in integrating into economic models a representation that makes public decision-makers react endogenously to events rather than behave exogenously.

The second development was that the failures of government intervention in areas such as macroeconomic management, employment, or development prompted research on political behavior. While some of these failures could be ascribed to genuine policy mistakes, insufficient knowledge, or simply bad luck, in other cases there was a need to provide explanations for a persistent inability to learn from past mistakes and from international experience. Why are certain regulations maintained, even though they obviously lead to outcomes that contradict stated policy objectives? Why had many developed countries returned to full employment by the 2000s while others were still experiencing mass unemployment? Why did some emerging market economies repeat in the 2000s the same errors (such as piling up foreign-currency-denominated and short-maturity debt) that had been made in the 1990s? If this were simply a matter of identifying appropriate policies, some form of learning should be at work and less-successful governments could be expected to learn from successful ones. Since some do not, clearly there is a need for political economy explanations.

The choice of a regime regarding product, capital, and labor market regulations involves *preferences* and tradeoffs between, say, efficiency and equity; economic *interests,* which can differ between, say, incumbents and newcomers; and *representations* of how the economy works, on which various players may disagree.[9] From a knowledge perspective, it is therefore important to understand these disagreements, to identify the economic interests involved, and to clarify the nature of the disagreements. From a policy perspective, recognizing and explicitly taking into account the intellectual and political environment of public decisions becomes as necessary as determining what is the first-best solution. Political economy then becomes essential both from a positive point of view (to understand why economic policy does not achieve its objectives) and from a normative one (to evaluate the chances of success of various reform strategies).

Positive economics, normative economics, and political economics thus coexist, and the modern approach of economic policy draws on all three methods. Positive economics remains indispensable to the understanding of the likely effects of public decisions. Normative economics brings intellectual

9. We return in more detail to the nature of those controversies in chapter 2.

discipline to policy choices and helps address the trade-offs they involve. Both, however, are nowadays aware of their own limits. They are increasingly supplemented by political economics.

Avinash Dixit (1996) once observed that the traditional approach of economic policy envisaged the ultimate policymaker—the Prince—as an omniscient, omnipotent, and benevolent dictator. The economics of imperfect information taught us that he or she was not omniscient. Second-best theory was developed in recognition of the fact that he or she was not omnipotent. Political economy tells us that he or she is not always benevolent. This should not be a cause for policy nihilism—only a motive against policy naiveté.

1.1.2 What do policymakers do?

Economic textbooks generally expand at length on economic structure and behavior but they tend to represent policymaking in a very sketchy way. They frequently assume that a single agent—the government—has sovereignty to decide to increase spending, cut taxes, raise the interest rate, manipulate the exchange rate, or introduce a minimum wage.

The actual situation is far from this caricature. Anyone sitting for a moment in the office of a finance minister can observe how diverse his or her responsibilities are, and how little time is actually devoted to making strategic decisions.

The main tasks of economic policymakers can be grouped into six categories:

1. *Set and enforce the rules of the economic game.* Economic legislation provides the framework for the decisions of private agents. Enforcement covers competition policy and the supervision of regulated markets such as banking and insurance. Economic legislation increasingly has an international dimension (through international treaties and agreements)—especially, but not only, in the European Union.
2. *Tax and spend.* Government spending amounts to about one-half of GDP in European countries and one-third in the UK, the US, and Japan. Budgetary decisions affect households' and firms' income and behavior through taxation and social insurance; they affect productivity through infrastructure, research, and education spending; and aggregate demand through changes in spending or overall taxation.
3. *Issue and manage the currency.* The choice of a monetary and exchange-rate regime is one of the most important single decisions a government can make. Defining and implementing monetary policy is the function of the central bank, which is responsible for setting interest rates, maintaining the value of the currency, and insuring that the banking system does not fall short of liquidity, even in the case of a crisis.

4. *Produce goods and services.* This is much less a government responsibility today than it used to be in the first decades after World War II, but most governments are still responsible for providing health care or education services, and some still own public enterprises in sectors like transport or energy.
5. *Fix problems or pretend to.* Ministers are frequently held responsible for a vast array of issues, from financial market turmoil to wage negotiations, company mergers, and plant closures and relocations. Many problems are beyond their means, but they can still try to influence private decisions—or at least pretend to.
6. *Negotiate with other countries.* Governments negotiate with other countries on trade liberalization and the definition of global rules. They participate in the governance of global and regional institutions (such as the International Monetary Fund, the World Trade Organization, or the European Union). They also participate in informal fora (G7, G8, G20, and regional summits) to hold discussions on global problems such as development, global warming, etc.

In fact, economic policy means different things to different people. In the US, the bulk of the policy discussion evolves around setting interest rates by the Federal Reserve Board and discussion in Congress on the President's tax and budget plans, and a limited set of specific issues such as energy security or healthcare reform. In Western Europe, the so-called *structural reforms**—i.e., attempts at changing labor market institutions, competition in product markets, health care insurance, and pensions—have taken center stage. In the last two decades or so, economic policy in Eastern Europe, China, and other transition economies has meant the introduction of markets and the privatization of state-owned companies. Finally, Argentina, Brazil, Turkey, and others have gone through long phases in which the sole obsession of policymakers was to control inflation and prevent—or cure—financial crises.

Economic policy also means different things in different times. Before the crisis that erupted in 2007–08 no policymaker thought she would have to design and implement a wholesale bank rescue, a large-scale budgetary stimulus or a massive expansion of the central banks' balance sheet.

To speak of "economic policy" in general may thus be regarded as presumptuous. However, there are many common features of economic policymaking across various contexts, fields, institutional setups, and time horizons, and they can be apprehended through a simple unified framework.

a) A simple representation of economic policy

We start by distinguishing objectives, instruments, and institutions.

- The *objectives** of economic policy are numerous (and sometimes contradictory): improving the population's standard of living, achieving full employment, maintaining price stability, reaching a fair distribution

of income, alleviating poverty, etc. They are sometimes explicitly stated in official texts. For example, the US "Full Employment and Balanced Growth Act" of 1978—known as the Humphrey–Hawkins Act—mandates the federal government to "promote full employment and production, increased real income, balanced growth, a balanced Federal budget, adequate productivity growth, proper attention to national priorities, achievement of an improved trade balance [. . .] and reasonable price stability." In the EU, Article 3 of the treaty on the European Union[10] states that the EU "shall work for the sustainable development of Europe based on balanced economic growth and price stability, a highly competitive social market economy, aiming at full employment and social progress, and a high level of protection and improvement of the quality of the environment [. . .] It shall combat social exclusion and discrimination, and shall promote social justice and protection, equality between women and men, solidarity between generations and protection of the rights of the child. It shall promote economic, social and territorial cohesion, and solidarity among Member States". What is immediately clear from such laundry lists of wishes is that economic policy has more than one objective and is easily given ambitious targets, irrespective of the difficulty or even impossibility of reaching all of them simultaneously.

- As already discussed, *instruments** are also numerous. Traditional ones relate to monetary policy (the setting of official interest rates) and fiscal policy (the choice of the levels of public expenditure and taxes). Economic policy is sometimes presented as a combination of these two instruments only. However, beyond them, it can and must rely on a variety of microeconomic instruments: regulations (from the provisions governing contracts and bankruptcy to sector-specific legislation), direct and indirect taxes on households and companies, subsidies, social security transfers, and even case-by-case decisions, as for competition policy.
- Lastly, *institutions** affect directly market equilibriums and the effectiveness of policy instruments. According to economic historian Douglass North (1993), "Institutions are the humanly devised constraints that structure human interaction. They are made up of formal constraints (rules, laws, constitutions), informal constraints (norms of behavior, conventions, and self imposed codes of conduct),

10. Frequently referred to as the "treaty of Rome," the "Maastricht treaty," or the "Lisbon treaty," the treaty establishing the European Community was signed in Rome in 1957 and amended several times, most significantly in Maastricht in 1991 to prepare for economic and monetary union, and in Lisbon in 2007. In what follows, we shall refer to it as the "EU Treaty," or sometimes as the "Maastricht Treaty" when referring specifically to its economic and monetary provisions. Note that the EU treaty in facts consists of two different documents, the "treaty on the European Union," and the "treaty on the functioning of the European Union." Concrete economic and monetary provisions belong to the latter.

arithmetic was successfully applied in the 1990s when a large number of central banks around the world became independent and inflation rates decreased dramatically (see chapter 4). The algebra of economic-policy decision-making is complicated, however, when the transmission mechanism is not known with certainty. Then, as shown by William Brainard (1967), the optimal policy setting should take into account the correlation between the parameters of the transmission mechanism and the objective variable. This creates a case for using several instruments to achieve a single target, as smaller movements in several instruments create less uncertainty than a large movement in a single instrument.

However, governments generally have many objectives but only a limited number of instruments. Hence, trade-offs are part of the governments' everyday life. Knowing trade-offs, choices are conditional on their preferences (for instance, how much more wage inequality they stand ready to accept to reduce the unemployment rate by one percentage point).

In such a setting, divergences in policy prescriptions can be either of a positive or of a normative nature: they can result from different views on the functioning of the economy (the constraint) or from different preferences, as represented by the loss function. It should also be noted that the loss function may itself have the character of an institution. For example, US law mandates the Federal Reserve System to try to achieve both price stability and full employment, while the EU Treaty assigns the European Central Bank price stability as an overriding objective (see chapter 4).

Such a representation was widely used in the 1960s. For instance, A.W. Phillips (1958) showed a negative relationship between the unemployment rate and the growth rate of nominal wages for the UK from 1861 to 1957. More specifically, he found that, with a 5% unemployment rate, wages were constant on average; with an unemployment rate slightly below 2.5%, wages increased by around 2% per year (Figure 1.1). This downward-sloping *Phillips curve** led to the idea of a trade-off between unemployment and inflation, a one percentage point fall in the unemployment rate having to be "paid back" by a rise in the inflation rate (here by 0.8 percentage points).

The responsibility of the economist was then to highlight and quantify this trade-off, that of the policymaker was to choose an inflation–unemployment combination according to collective preferences. As developed in this book, the simultaneous rise of inflation and unemployment in the 1970s challenged this excessively simple representation.

c) Changing the institutions: structural reform

The trade-offs just described are generally reversible: the central bank raises or cuts the interest rate according to the economic situation, parliament increases or reduces taxes, etc. However, in the 1980s and 1990s, persistent problems in growth and employment in Europe highlighted the limits of such economic management. A good example here is the apparent trade-off

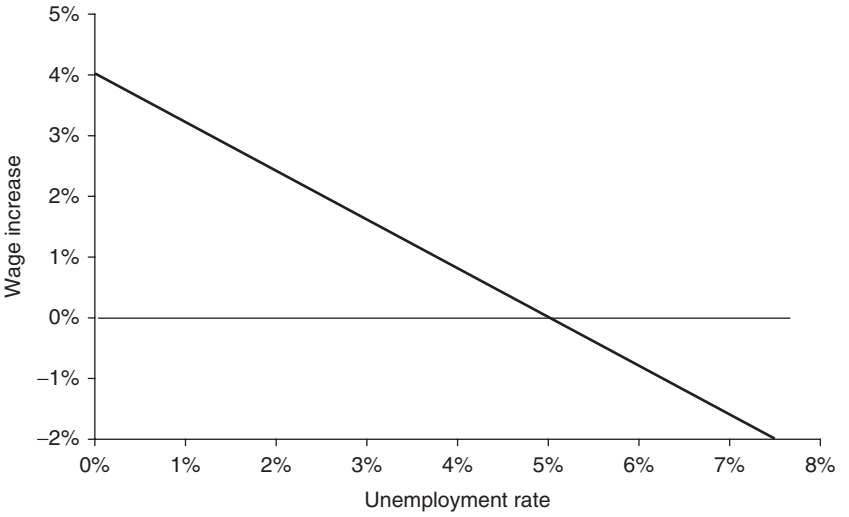


Figure 1.1 The curve of A.W. Phillips.
Source: Phillips (1958).
Note: For the sake of simplicity, the Phillips relationship has been linearized.

Box 1.1 Trade-offs and Economic Management

A government has n target variables $Y_1, Y_2, \ldots Y_n$ represented by a vector $Y = (Y_1, Y_2, \ldots Y_n)$, and n corresponding objectives. Its preferences can be summarized by a loss function L that measures the welfare loss associated with a divergence between the value taken by the target variables Y_i and their objective values $\tilde{Y}_i$:

$$L(Y_1 - \tilde{Y}_1, Y_2 - \tilde{Y}_2, \ldots Y_n - \tilde{Y}_n) \qquad \text{(B1.1.1)}$$

L is a convex, continuously differentiable function with $L(0, 0, \ldots 0) = 0$. There are p independent policy instruments that can be grouped in a p-dimensional vector $X = (X_1, X_2 \ldots X_p)$. With I representing the institutions, the functioning of the economy can be represented by:

$$Y = H_I(X) \qquad \text{(B1.1.2)}$$

Economic policy then consists in selecting X such that L is minimized, conditional on (B1.1.2).

If $n = p$, then it is usually possible to invert (B1.1.2) and find the vector X which allows Y to be exactly at its target level.

If $n > p$, this is no longer the case, and the government faces a trade-off. In other words, the program leads to choosing values for $(X_1, X_2 \ldots X_p)$ such that, at the margin, it is not possible to improve on any of the targets

> and their enforcement characteristics. Together they define the incentive structure of societies and specifically economies." Lasting features of the organization of products, labor, and capital markets (i.e., the bankruptcy code, the rules governing employment contracts, the legislation on takeovers) or of the framework for economic policy decisions (i.e., budgetary procedures, the statute of the central bank, the exchange-rate regime, the rules governing competition, etc.) are regarded as institutions. This definition includes nonpublic institutions such as, for example, trade unions, which are private associations but affect the functioning of labor markets.

Within this framework, institutions represent a kind of social capital. They are not eternal and can evolve, be reformed, or disappear, but they have some permanence and can be taken as given for the traditional analysis of policy choices.

b) Economic policy as a succession of trade-offs

Consider a government that targets n different economic variables, such as the unemployment rate, the inflation rate, and the current account (in this case, $n = 3$), and has a specific objective for each of them. For instance, the government wants an unemployment rate of around 5% of the active population, an inflation rate of around 2% per year, and a balanced current account. The preferences of the government can be summarized by a *loss function** that depends on the difference between each target variable and its desired value.

Assume now that the government has p independent policy instruments, i.e., p variables that it can handle directly (for instance, the fiscal balance and the short-term interest rate, in which case $p = 2$). Economic policy then consists in setting the p policy variables such that the loss function is minimized.

If $p = n$, then the n policy objectives can all be achieved, because there is an equal number of instruments (see box 1.1). In our example, however, we have $p < n$ and the n objectives cannot be achieved simultaneously, which implies trading off one objective against another one. For instance, the government needs to accept a current-account deficit if it wants to lower unemployment to a level close to 5% while keeping inflation close to 2%. More generally, to reach n independent policy objectives, the government needs at least an equal number of policy instruments. This is known as the *Tinbergen rule**.[11]

One direct implication of the Tinbergen rule is that an independent central bank with a single objective of price stability will be able to reach this objective since it can fully make use of one instrument (monetary policy). This piece of

11. After Jan Tinbergen, the Dutch economist who was awarded the first Nobel Prize in Economics in 1969 for his work on economic policy (Tinbergen, 1952).

without welfare deteriorating due to a higher divergence on other targets. Analytically, this corresponds to a situation where:

$$dL = \sum_{i=1}^{n} \frac{\partial L}{\partial Y_i} dY_i = 0$$

i.e., for any pair (i, j) of objective variables,

$$\frac{dY_i}{dY_j} = -\frac{\partial L/\partial Y_j}{\partial L/\partial Y_i}$$

The *marginal rate of substitution** between any two objectives is therefore equal to the inverse ratio of the partial derivatives of the loss function. This formula, formally identical to what is obtained in a consumption maximization program, means that at the minimum of the loss function, any improvement in an objective is compensated by a deterioration in another one in inverse proportion to the effects of these variations on the loss function.

between employment and productivity. In some European countries fewer people work, but those who work have a high level of productivity. Other countries achieve much better performances as regards employment, but at the price of weaker productivity. Collectively, the European countries seem confronted with a trade-off described by the negatively sloped AA curve of figure 1.2.

Attempts at modifying the position of a country along the AA schedule through various levers such as tax rates and public spending can be characterized as *economic management*.

However, trading off more jobs for less income per worker is unsatisfactory. In a low-employment situation the true objective of economic policy should be to reach *at the same time* higher employment and higher productivity levels. The right answer would therefore consist in moving AA outward, thereby simultaneously raising employment and productivity. This requires reshaping institutions: For example, stronger incentives to remain active and take up jobs, more investment in education, an environment that fosters innovation, etc.

In a more general way, structural reforms aim at modifying economic policy trade-offs by changing the institutions. A study by the International Monetary Fund (2004) defines them as entailing "measures that, broadly speaking, change the institutional framework and constraints governing market behavior and outcomes." To see what this means, let us take the simple case where there are two objective variables Y_1 and Y_2, with only one instrument X to reach them, and, therefore:

$$Y_1 = h_I^1(X), \quad Y_2 = h_I^2(X) \tag{1.1}$$

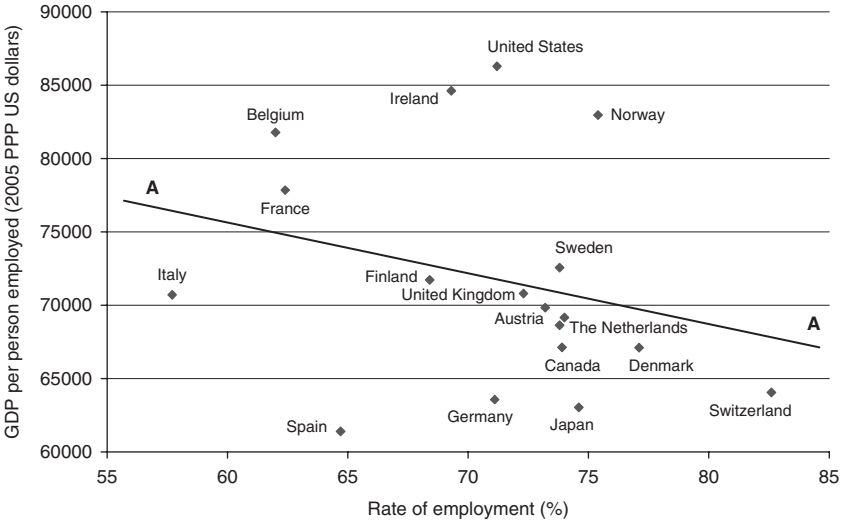


Figure 1.2 The employment–productivity trade-off in 2005.
Source: Authors' calculations using Groningen Growth and Development Center and OECD data.

where I represents the institutions. The instrument X can be substituted in the two relations, giving an explicit formulation of the trade-off between Y_1 and Y_2, conditional on the institutions:

$$g_I(Y_1, Y_2) = 0 \tag{1.2}$$

Structural reform aims at substituting institutions J for institutions I to improve the trade-off between Y_1 and Y_2 (figure 1.3).

It is common, but inaccurate, to assimilate structural and supply-side policies. Making the central bank independent, choosing a new currency regime, or adopting a framework for budgetary policy are true structural reforms because they aim at improving existing trade-offs between various objectives by moving the corresponding schedules outward. Contrarily, a change in tax rates, which is mostly a supply-side measure, does not have the character of a structural reform.

However, many of the structural reforms undertaken since the 1980s in advanced economies were admittedly of a supply-side nature. Widespread reform of capital markets through the elimination of credit controls, the scrapping of many deposit regulations, and the liberalization of capital flows had major consequences, both micro- and macroeconomic. Deregulation in product markets, following its initiation in the US in the 1970s, increased competition and fostered innovation, resulting in productivity gains, especially in sectors such as transport, telecommunications, and energy. In the EU, the

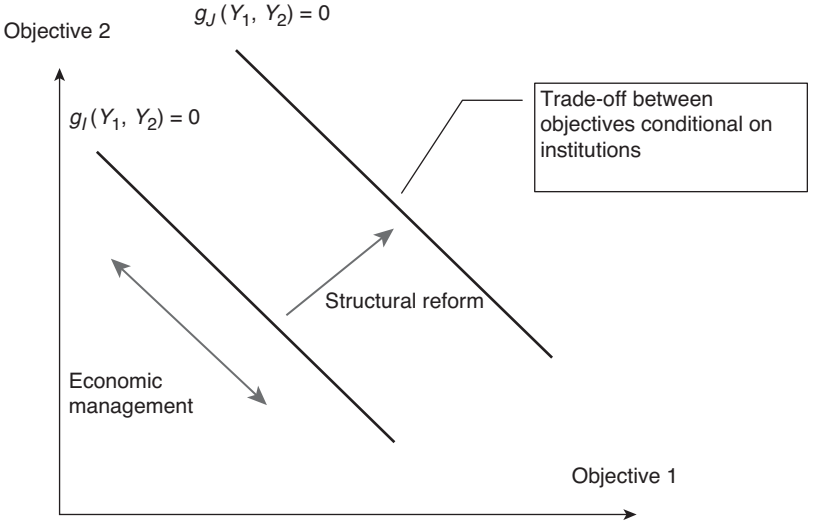


Figure 1.3 From managing trade-offs to reforming institutions: An illustration.

gradual introduction starting in the mid-1980s of a *single market**[12] for goods and, to a lesser extent for services, had similar objectives. In developing and emerging countries, the standard concept is that of *structural adjustment**—a package of reforms advocated by the International Monetary Fund and the World Bank and enforced on countries requiring financial assistance. Though somewhat broader, structural adjustment encompasses several features of what we call structural reform.

Structural reforms are often viewed as having negative short-term, but positive long-term effects. The most telling example of such effects was, at the end of the twentieth century, the transition of the former planned economies of Central and Eastern Europe and the former USSR to market economies. Figure 1.4 highlights the GDP cost of this transformation: It generally took several years before GDP returned to its pre-transition level. Furthermore, some of the most successful post-transition countries where those, like the Baltic States, where the initial fall was the most pronounced. While less dramatic, many structural reforms have the character of an investment whose costs are paid up-front while it yields benefits only over the medium run. This, for example, was the case with the disinflation and exchange-rate stability policies pursued in Europe in the 1980s and the 1990s.

12. Within a single market, not only are the customs duties eliminated, but products and factors of production (capital and workers) move without obstacles. Also, there are no obstacles to the cross-border provision of services.

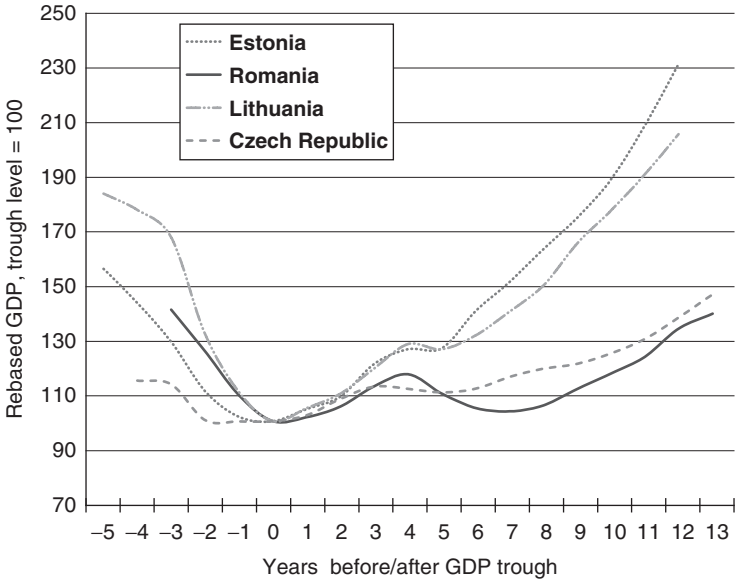


Figure 1.4 GDP impact of the transition to the market economy.
Source: Authors' calculations based on the Groningen Growth and Development Center's Global Economic database.

Such intertemporal effects necessarily raise political economy issues. For a democratic government facing a reelection constraint, undertaking reforms that will antagonize voters and only yield benefits after its term expires can be a recipe for failure. How to surmount this political economy constraint (for example by finding ways to compensate incumbents for the rents they will lose as a result of the reform) is a major theme for research.

1.2 The Whys and Hows of Public Intervention

Having presented what policymakers do and how economic policy works, let us move to an upstream question: why is public intervention needed? What are the objectives of public intervention? To this rather naïve question, economic theory provides rather precise answers.

1.2.1 The three functions of economic policy

Musgrave and Musgrave (1989) have distinguished three essential functions of budgetary and, more largely, economic policy:

- *Allocation** of resources (i.e., their assignment to alternative uses). This covers public interventions aiming at affecting the quantity or the

quality of the factors (capital, unskilled and skilled labor, technology, land, etc.) available for production, and their sectoral or regional distribution. More generally, policies aiming at the provision of public goods such as infrastructure building or environmental preservation are included in this category.

- *Macroeconomic stabilization** vis-à-vis exogenous shocks that move the economy away from internal balance (defined as full employment together with price stability). This covers policies aiming at bringing the economy closer to balance—a role that Keynesian economists usually assign to monetary and budgetary policies.
- *Income redistribution** between agents or regions. This covers policies aiming at correcting the primary distribution of income. Progressive taxation policies and social transfers are key instruments to this end.

Redistribution has a different scope than either allocation or stabilization since it addresses the distribution of income within society. However, allocation and stabilization may seem to pursue similar goals. The distinction between them directly refers to the distinction between long-term output growth and short-term fluctuations around the trend: allocation policies aim at increasing the maximum level of output that can be reached without creating inflation—what is generally called *potential output**, while stabilization policies aim at minimizing the divergence between actual and potential output, known as the *output gap** (figure 1.5 and box 1.2).

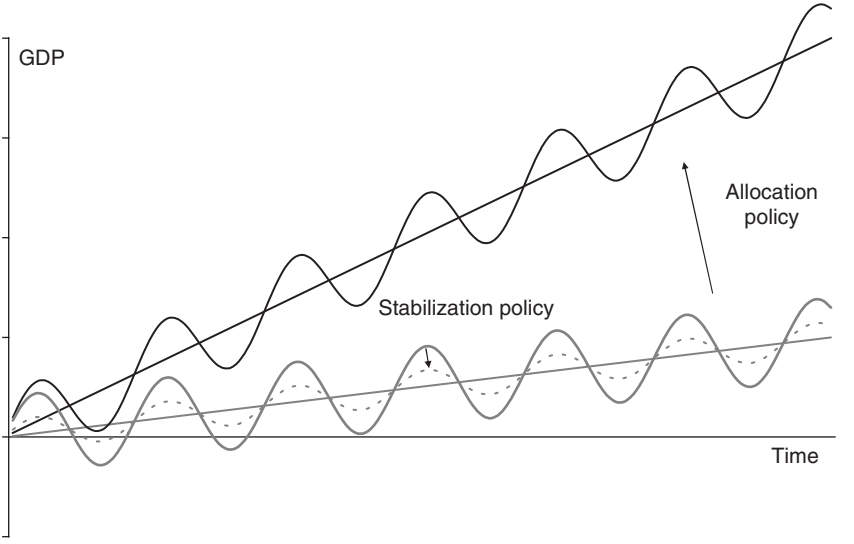


Figure 1.5 Stabilization versus allocation policies.

Box 1.2 Supply, Demand, and the Output Gap

In a simple model of the supply side of the economy, potential output is determined by the factors of production (mainly labor supply and the capital stock), as well as by the factors affecting productive efficiency. A standard representation is:

$$Y_t = F_t(K_t, N_t) \tag{B1.2.1}$$

where Y is production, K the capital stock, N employment, and F the production function. K and N depend on time, and so does F as improvements in technology allow more to be produced with the same amount of factors.

In the short run K can be considered exogenous, so $K_t = \overline{K}_t$. Let us define $\overline{N}_t$ as the employment level that is reached when unemployment rate is at a level $\bar{u}_t$ called the *equilibrium rate of unemployment**. $\bar{u}_t$ cannot be zero because at each point in time, a fraction of the labor force is looking for a job. Its level depends on the efficiency of the country's labor market institutions. So if $\bar{L}_t$ is the labor force,

$$\overline{N}_t = (1 - \bar{u}_t)\overline{L}_t \tag{B1.2.2}$$

Potential output can thus be defined as:

$$\overline{Y}_t = F_t(\overline{K}_t, \overline{N}_t) \tag{B1.2.3}$$

It is exogenous in the short term but endogenous in the long term as the capital stock adjusts.

The *output gap** can thus be defined as the difference between the demand-determined output Y_t and the supply-determined potential output Y_t. It is generally measured as a percentage of the potential output, so:

$$output\ gap = \frac{Y_t}{\overline{Y}_t} - 1 \tag{B1.2.4}$$

A negative output gap means that production is below potential, implying non-equilibrium (or involuntary) unemployment. A positive output gap means that production is above potential. This may look strange if one thinks of the capital stock and the available labor force as a physical constraint. However, there are ways to adjust to a higher level of demand. For example, a standard response to excess demand is to have recourse to overtime; or older equipment that was regarded as obsolete but had not been discarded can also be put to use again; or less-efficient producers, who were hardly able to compete in normal conditions, may increase their supply. However, such responses tend to be costly, implying

a rise in the marginal cost of production and therefore a rise in aggregate price level.

The output gap is a simple notion but it is hard to measure in practice, because the capital stock $\overline{K}_t$, the equilibrium rate of unemployment $\bar{u}_t$, and the production function F are all unobservable (this is less true for the capital stock that could be measured through surveys, but in practice it is generally evaluated on the basis of past investments and assumptions as regards the annual rate of discard). The various available measures, such as those provided by international institutions (such as the IMF, the OECD, and the European Commission) differ significantly and are frequently revised. Because of these difficulties, potential output is sometimes derived from actual output through purely statistical techniques (by applying a filter to the actual series to estimate its trend). However, this ignores the fact that potential output is an economic notion and that its level depends on prices: for example, a higher price of energy reduces potential output because it makes certain energy-intensive production techniques unprofitable. Statistical shortcuts are therefore inappropriate in the presence of economic shocks.

This makes it difficult to base policy choices on an accurate evaluation of the output gap. This especially applies to countries whose trend growth rate has not remained constant over time. A comparison between the US and France (or other countries that have gone through a catching-up period during which their growth rate has increased) is telling in this respect. Prior to the global crisis, the US growth rate did fluctuate but the trend was roughly stable: a simple linear trend over a long period of time captured most of the long-term evolution (figure B1.2.1.a). The French case was very different, as the trend in growth rate had decreased from about 5% in the 1960s to less than 2% in the mid-2000s (figure B1.2.1.b). This implies that a French policymaker observing the evolution of GDP in real time could have mistakenly diagnosed a negative output gap while it was in fact the growth-rate trend that was slowing down (in fact, this is what happened in the 1970s and again in the 1980s).

The issue became acute in the aftermath of the global recession of 2008–09: Policymakers were at pains to determine to what degree the sharp output decline experienced by most countries could be recouped by future above-trend growth. Optimists were considering that the crisis had mainly affected demand, not potential output, and that the output gap was therefore very large. Pessimists were objecting that foregone capital accumulation, the withdrawal from the labor force of workers discouraged by the rise in unemployment, and the higher cost of credit all resulted in a lowering of potential output. This debate, to which we return in chapter 8, has profound consequences for monetary and budgetary policies.

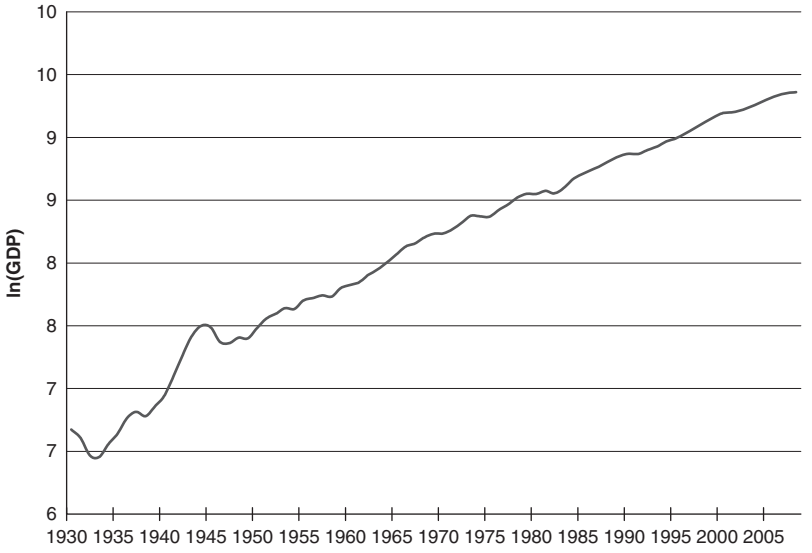


Figure B1.2.1a US real GDP, 1930–2008.
Source: US Bureau of Economic Analysis, and authors' own calculations.

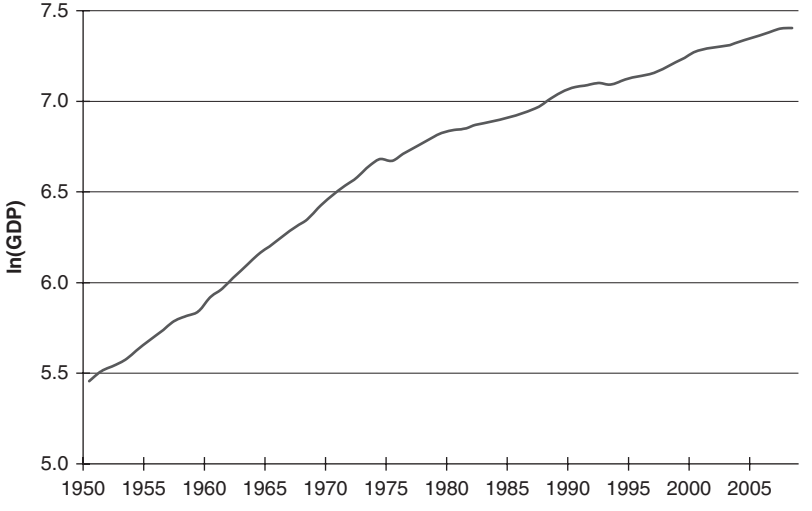


Figure B1.2.1b French real GDP, 1950–2008.
Source: Institut National de la Statistique et des Etudes Economiques (INSEE), and authors' own calculations.

This distinction between three main functions is widely used in policy discussions; it helps bring some discipline and helps clarify the aims of policy decisions. The distinction is followed in this book, of which chapters 3–5 deal primarily with stabilization, chapters 6 and 7 with allocation, and chapter 6, also with redistribution. As we will see, however, there are many reasons why these three functions frequently interfere with each other, making economic policy choices less clear-cut than in this simple presentation.

1.2.2 Why intervene?

For economists, public intervention requires justification. This is because the *first theorem of welfare economics** establishes that any competitive equilibrium is a *Pareto optimum**. In other words, it is not possible to improve the welfare of an economic agent without reducing that of another one.

This is both a very powerful and a very limited result. It is powerful because if public intervention can improve the fate of some agents only by deteriorating that of others, this immediately raises the question of the moral basis and the acceptability of such an intervention. It is, however, limited for two reasons. First, the Pareto criterion is silent on the distribution of income and wealth between economic agents (any distribution can be considered Pareto-optimal). Second, the conditions for this result to hold are very strict ones: Kenneth Arrow and Gérard Debreu (1954) have shown that the first theorem of welfare economics relies on a very demanding set of hypotheses. A true competitive equilibrium requires, inter alia, strictly competitive markets, the existence of a complete set of markets that allows the carrying out of transactions on all goods at all periods, and perfect information. Challenge one of these hypotheses, and there is justification for public intervention.

In fact, this welfare theorem, which was often interpreted as providing the doctrinal basis of *laissez-faire*, can just as well provide arguments for the partisans of public intervention, provided they are able to justify it by precise arguments.

a) Allocation

As concerns allocation, arguments are microeconomic in nature. State intervention is justified when it is able to remedy market failures, i.e., to improve the efficiency of resource allocation in comparison to the market outcome. The most frequent reasons for such failures are the presence of monopolies, *externalities**,[13] the existence of public goods, information asymmetries between agents, market incompleteness, or agent shortsightedness.

13. An *externality**, also called an external effect or a spillover, is the (positive or negative) effect of production or consumption on agents who were not participants in the decision to produce or consume. For example, motor vehicle transportation creates negative externalities through road congestion, noise, and pollution. Positive externalities can, for example, be found in the use of a

These arguments, which have been extensively studied in microeconomics and public economics, traditionally provide solid justifications for regulatory policies, corrective taxation, the public provision of certain goods and services, or public subsidies (box 1.3).

Box 1.3 Microeconomic Arguments for Public Intervention

Public intervention has justification when one of the hypotheses of the first theorem of welfare economics is violated.

Competition Is Not Perfect

Rationale. Profit maximization by a company implies equalizing the marginal cost (of producing an additional unit) and the marginal income (from selling an additional unit). Under perfect competition, the marginal income is the market price of the product and profit maximization leads to a social optimum. If a firm holds a monopoly position or, more generally, has some *market power**,[14] it takes into account the (less than infinite) elasticity of demand for its product and the fact that its marginal income is less than the market price. This is because selling an additional unit implies lowering the price of all previous units. In comparison with the perfect competition outcome, this leads the firm to reduce quantities sold and to increase the price, to the detriment of the consumer.

Public intervention can aim at restoring perfect competition conditions (for example, by blocking mergers leading to, or threatening to lead to, excessive market power). However, it is not always desirable to eliminate monopolies: when production involves high fixed costs or, in general, when there are increasing *returns to scale**,[15] larger firms or even monopolies are more efficient than smaller ones. This is what is meant by *natural monopoly**. For example, it is more efficient to have the railway network managed by a single entity than by several, but this implies regulating its behavior or subjecting it to potential competition (via granting it a fixed-duration contract only) in order to prevent it from exploiting its monopoly power.

network-based software such as eBay, Skype, or Facebook: Its usefulness for any user increases with the number of users connected.

14. Market power is the possibility for a producer to set a price higher than its marginal production cost (the equilibrium price on a competitive market). This happens when competition is not perfect and the demand for a firm's product is less than infinitely elastic.

15. The return to scale measures the relative increase in production resulting from an increase in the volume of all factors of production (capital, labor, etc.) by a factor k. Returns to scale are increasing if production increases by more than k, decreasing if it increases by less than k, constant if it increases by k.

Consequences. This argument first and foremost constitutes the prime justification of competition policy, whose aim is to prevent firms from acquiring a dominant position or from abusing it. In the name of consumer protection, institutions in charge of it, such as the Federal Trade Commission in the US, the Bundeskartellamt in Germany or the European Commission in the EU, can block mergers and acquisitions if they threaten to create monopolies, or fine companies that abuse a dominant position in their market. For example the European Commission (which is in charge of this policy for cases with a cross-border dimension), blocked several merger operations (Alcan–Alusuisse–Péchiney in 2000; GE–Honeywell in 2001) that were regarded as a potential threat to competition. It also levied fines on companies whose behavior was regarded as obstructing competition. In 2004, Microsoft was fined €497 million for abusing its market power in the EU. However, the argument has wider applications: For example, it provides a justification for setting a minimum wage if employers locally hold near-monopoly positions as purchasers of unskilled labor (this is called a monopsony).

Economic Activities Have External Effects

Rationale. In the presence of externalities, the private cost of a resource or the private profit from production do not coincide with the social cost or the social benefit. For example, this can be the case for a firm which consumes a natural resource such as clean water, or whose production technique spoils the environment, but which does not take the corresponding social costs into account in its profit maximization. In such cases, the firm tends to over-consume natural resources and to overproduce. The reverse occurs when the externality is positive (i.e., if production has favorable nonmarketable effects). For example, a research-and-development-intensive firm that establishes a facility in an area tends to exert positive effects on other firms through the development of local suppliers and subcontractors, the creation of a more liquid market for skilled labor, and links with university departments. However, those positive externalities are not taken into account in the decision by the firm to open a new facility, which leads to a sub-optimal number of such facilities. It is also the high negative externalities from the default of large financial institutions that justify rescuing banks in a financial crisis. The risk is that a bank default would make other financial institutions insolvent, thereby triggering a chain reaction.

Consequences. Environmental economics largely rests on this type of argument, both as regards local pollution (water and air spoilage, waste, etc.) and global pollution (greenhouse effect). The first-best economic response (not necessarily the most frequent one) generally consists in letting agents "internalize" externalities by taxing the negative ones (this is the so-called

polluter–payer principle in use in several countries) and by subsidizing the positive ones (local governments routinely subsidize investment from nonresident companies or grant them tax exemptions). However, here again, the argument is broader: A company which lays off its employees exerts a negative externality on the community, which bears the cost of unemployment insurance, and the one that hires creates a positive externality. This justifies making a company's contributions to unemployment insurance a function of its hiring and firing behavior, as is the case in the US. Olivier Blanchard and Jean Tirole (2008) have proposed extending such experience rating to Europe. As regards the risks of letting a major financial institution default on its liabilities, the dramatic consequence of the Lehman Brothers bankruptcy in 2008 and the rescue of a series of other US and European banks in the following months illustrate the importance of state intervention. We return to this discussion in chapters 4 and 8.

Information Is Imperfect

Rationale. The optimality of the competitive equilibrium rests on a perfect information hypothesis. If information has a strategic character and if agents use it to their profit, the market outcome is no longer necessarily Pareto-optimum. The potency of this argument was recognized with the awarding of the 2001 Nobel Prize to George Akerlof and Joseph Stiglitz, who contributed to the development of the economics of imperfect information. Stiglitz and Weiss (1981), for example, showed that when the creditor (say, a bank) has less information than the debtor (say, a company) on the risk incurred in lending, it cannot accurately price the risk in setting the interest rate on the loan. To prevent the pricing of credit without regard to debtor-specific risk resulting in selecting the riskiest borrowers (a phenomenon known as *adverse selection*[*16]), it is optimum for the creditor to ration credit, which is socially inefficient (see chapter 4).

Consequences. Imperfect information is pervasive, but it also affects policymakers, who rarely enjoy an undisputed informational advantage. Public policy can foster the dissemination of market-relevant information, either in the form of aggregate statistics (the IMF was given an enhanced role in this respect after the emerging-countries financial crises of the 1990s,

16. Adverse selection takes place when information asymmetry leads to elimination of the most efficient suppliers or buyers from the market. The standard example is that of the market for second-hand cars described by Akerlof (1970): Only the sellers know the quality of the vehicles they sell. The competitive selling price corresponds to average quality; therefore, sellers of high-quality vehicles find the price too low and reject selling their car. The result is a fall of average quality, and therefore of the price. Eventually, only the lemons may be put on sale. Such adverse selection is obviously not optimal. This problem is common in the insurance business.

through the so-called "Special Data Dissemination Standard") or through standardizing the publication of company-specific information. Accounting and financial reporting standards, for example, are intended to ensure that financial markets benefit from comparable, undistorted information. As illustrated by the Enron affair, this is by no means an easy task: In particular, the accounts published by the same company can differ under competing reporting standards (e.g., the International Financial Reporting Standards (IFRS) used in Europe and the Generally Agreed Accounting Principles (GAAP) of the US). Beyond such focus on transparency, imperfect information, for example, justifies the regulation of certain businesses such as insurance, as well as government support for financing newly created companies.

Markets Are Incomplete

Rationale. The optimality of the competitive market equilibrium hinges on the existence of markets for all necessary transactions at all relevant horizons. When such markets are missing, Pareto-optimality is not guaranteed. For example, borrowing to finance one's education is made difficult by the absence of collateral on which the loan can be guaranteed, and by the fact that the choice of a professional specialization is hardly reversible. The near-absence of a market on which young people could borrow to finance investment in their own human capital tends to limit access to higher education, especially in developing countries. In the absence of public intervention, private investment in human capital is therefore sub-optimal, which harms growth.

Applications. This argument provides a justification for government to step in where markets are missing. In the above example, it gives an economic efficiency motive for providing grants and scholarships to students or to ensure the public provision of education services. However, governments can also create new markets: In the 1990s, Australia and New Zealand have pioneered the introduction of income-contingent loans to students, the repayment of which is a function of the beneficiary's future income, and a number of other countries such as Chile, the UK, South Africa and Thailand have followed suit (Chapman, 2006). Such reforms are frequently introduced as a counterpart to an increase in tuition fees (this was the case in the UK in 2004). Robert Shiller (2003) has proposed to go beyond this and develop specific financial products in order to insure the students against the risk that economic change devalues their human capital. In another field, government debt agencies have introduced inflation-indexed bonds. Such instruments provide private agents with a way to hedge their fixed-income savings against the risk of future inflation.

b) Stabilization

While public intervention in the name of allocation aims at altering the long-run market equilibrium, intervention carried out in the name of stabilization is intended to limit short-term deviations from it. The motive remains the search for efficiency, but it is not the possible inefficiency of the equilibrium that matters, but rather the efficiency loss resulting from not reaching it.

Keynes gave two reasons for such intervention. The first one is what he called "animal spirits", the instability of private behavior under the influence of spontaneous expectations leading to excessive optimism followed by excesses of pessimism:

> Even apart from the instability due to speculation, there is the instability due to the characteristic of human nature that a large proportion of our positive activities depend on spontaneous optimism rather than on a mathematical expectation, whether moral or hedonistic or economic. Most, probably, of our decisions to do something positive, the full consequences of which will be drawn out over many days to come, can only be taken as a result of animal spirits—of a spontaneous urge to action rather than inaction, and not as the outcome of a weighted average of quantitative benefits multiplied by quantitative probabilities.
>
> Keynes (1936), chapter 12, paragraph 7

Second, Keynes argued that *nominal rigidities** [17] of wages and prices prevent the self-correcting market mechanisms from operating and moving the economy back to equilibrium. Especially, nominal wage rigidity implies that the *real wage** (i.e., the nominal wage divided by the price level, which is a measure of the real cost of labor) does not fall in a recession, preventing the restoration of full employment.

In the eyes of Keynes, the combination of private instability and ineffective self-correcting mechanisms provided a justification for relying on *counter-cyclical** monetary and fiscal policies to smooth out economic fluctuations and prevent economic depressions. As illustrated in figure 1.5, such stabilization policies are distinct from allocation policies, which aim at making the economy more efficient, and hence at raising the growth rate of the economy in the long run.

17. Rigidities designate a failure of prices or wages to adjust in response to changes in economic conditions. Economists usually distinguish *nominal rigidities* from *real rigidities**. The former arise from a stickiness in the setting of nominal prices. For example, the wages of employees holding jobs do not change when unemployment varies, or companies do not adjust their price lists when demand falls. Real rigidities are of the same nature but affect real variables such as the real wage, the real interest rate, etc. For example, nominal wages may change as well as the price of goods without their ratio (the real wage) being affected. Nominal rigidities frequently imply real rigidities, but there can be real rigidities in the absence of nominal rigidities.

The arguments for stabilization policies have since inception been a matter for theoretical and empirical disputes, especially from the 1970s to the late 1980s, the high noon of the monetarist backlash. Yet economic fluctuations remain a fact of life and accounting for them while remaining consistent with rational behavior assumptions has proved to be challenging. The theory of *real business cycles** developed in the 1980s was a conceptually coherent attempt at explaining fluctuations by *shocks** to the production technology and rational responses to them by maximizing agents—thus without relying in any significant way on irrational behavior or nominal rigidities. However, in spite of the considerable literature devoted to this approach, its empirical relevance for the explanation of short-term fluctuations remains disputed.[18]

Of the two explanations offered by Keynes, the first—the notion that economic agents are driven by "animal spirits" rather than by cool-headed rational calculation—was and remains in contradiction with the basic assumptions of economics. Though risk premiums in financial markets do vary over time, and in spite of recent developments in experimental economics, which indicate that departures from rational behavior are frequent, the animal spirits assumption remains alien to the methodological foundations of the economic profession. As emphasized by scholars of crises such as Kindelberger (1978) and Minsky (1992), and as observed in 2007–09, it has however relevance in situations of financial panic.

The argument based on nominal rigidities is theoretically closer to mainstream economics, provided an explanation is given for why and how such rigidities affect economic behavior. As developed in chapter 4, the standard response long remained the somewhat ad-hoc argument that agents enter into contractual arrangements that involve nominal rigidities—for example, wage contracts that specify a nominal compensation and are only renegotiated at discrete intervals. It was only in the 1980s that Keynesian economists provided convincing micro-founded explanations for nominal rigidities by showing that the gain to the microeconomic agent from changing prices in response to a shock can be much smaller than the corresponding macroeconomic benefit.

Where contemporary macroeconomics has been successful is in providing a framework for thinking about the role of stabilization policy, and for distinguishing between situations where it is effective and situations where it is ineffective.

This approach is based on a simple aggregate supply-and-demand framework that depicts the relationship between potential output and the product price, on the one hand, and between aggregate product demand and the product price on the other. In the short run, aggregate supply depends positively on the product price, as depicted by the aggregate supply curve,

18. The real business cycle literature originates in the work of Kydland and Prescott (1982). Galí and Rabanal (2004) provide a sceptical account of its empirical relevance to the US case.

because in the presence of nominal rigidities a rise in the price level reduces the real wage and makes production more profitable. In the long run, aggregate supply is fixed as unemployment is at its equilibrium level, so the curve is vertical. Aggregate demand depends negatively on it, as a rise in price reduces the real value of nominal assets and thereby reduces consumption. The two relationships are depicted by the aggregate supply and aggregate demand curves in figure 1.6 (see box 1.4 for a formal derivation).

In this context two distinctions need to be made. The first one is between variations of the quantity supplied or demanded in response to a change in the product price (a move along the supply-and-demand schedules in figure 1.6) and exogenous perturbations (movement of the whole schedules), interpreted as shocks to the economy. The second one is between shocks to supply and shocks to demand. *Supply shocks** and *demand shocks** have become part of every macroeconomic policymaker's toolkit:

- A supply shock is an exogenous modification in the relationship between potential output and the product price. For example, at any given level of the wage and the product price an oil shock (a rise in the price of oil) reduces the level of potential output because it increases prices and reduces the profitability of production.
- A demand shock is an exogenous modification in the relationship between product demand and the product price. This can be for example a drop in the level of household consumption resulting from a reduction of household wealth.

Although both kinds of shocks may result in a reduction or a rise in output, they command different policy responses and it is important to sort out one from the other. This can be understood through the formal representation of the balance between aggregate supply and aggregate demand represented in figure 1.6.

A positive demand shock shifts aggregate demand to the right, resulting in moving from the initial equilibrium E to A′, characterized by both a higher output and a higher price. A positive supply shock, however, shifts aggregate supply to the right, resulting also in higher output but a lower price (B′). So the simple criterion for distinguishing demand from supply shocks is that for a similar effect on output they result in opposite changes in the price.

In the long run, the aggregate supply curve becomes vertical because capital adjusts fully and unemployment is supposed to be at its equilibrium level. The reasoning is the same except that a positive demand shock now exclusively results in a price rise as the equilibrium moves from E to A″. For a supply shock, the result is qualitatively unchanged as the equilibrium moves from E to B″.

The upshot is that a demand shock either does not affect output or moves it in the same direction as price, while a supply shock either does not affect price or moves it in the opposite direction to that of output.

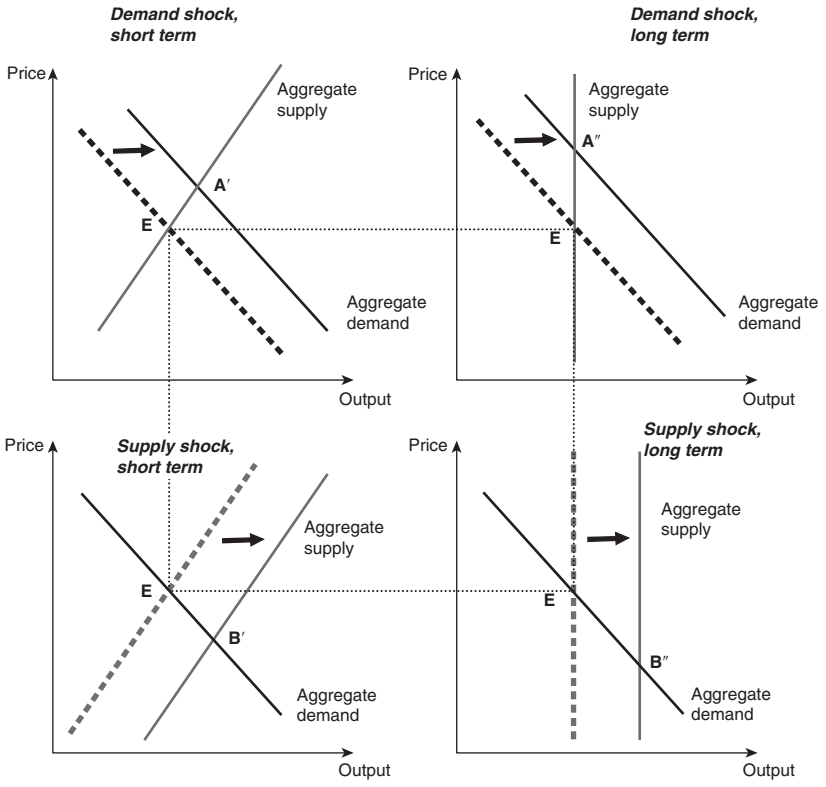


Figure 1.6 Supply and demand shocks in an aggregate supply–aggregate demand framework.

This framework helps understand the role and the limits of stabilization. A monetary or fiscal impulse affects the demand curve and can therefore offset a demand shock. This is for example the elementary reasoning behind the response of the Federal Reserve and of the US federal fiscal policy to the depression of the late 2000s: As household wealth declined, macroeconomic policy aimed at stimulating private demand through lower interest rates and direct transfers to households. However, a fiscal or monetary impulse does not affect the aggregate supply curve, so they are ineffective in response to a supply shock. If the supply curve shifts to the left in response to a rise in the relative price of oil (which makes other products less profitable and therefore reduces supply), pushing aggregate demand to the right necessarily results in a further increase in the price level and is fully ineffective in the long run. Therefore, demand policies are only effective in response to certain categories of shocks.

Box 1.4 Aggregate Supply and Aggregate Demand

Deriving the aggregate demand curve is straightforward. It is natural to suppose that household consumption depends positively on household real wealth. Real wealth in turn depends negatively on the price level as part of the assets, such as cash, bank deposits, and bonds, are denominated in nominal terms. Thus,

$$C = C\left(Y, \frac{\Omega}{P}\right) \text{ with } \frac{\partial C}{\partial Y} > 0, \frac{\partial C}{\partial\left(\frac{\Omega}{P}\right)} > 0 \text{ and therefore } \frac{dC}{dP} < 0 \tag{B1.4.1}$$

where C is household consumption, Y is household income, Ω household wealth, and P the product price level.

Constructing the supply curve is slightly more complex. Let us first suppose that labor is the only factor of production and that the quantity of it employed in production N is bound to $\overline{N}$ as in box 1.2. Suppose also that the marginal productivity of labor is decreasing, for example because employers start by hiring the best trained and most productive employees. Then,

$$Y = AN^{\alpha} \text{ with } 0 < \alpha < 1 \text{ and } A > 0 \tag{B1.4.2}$$

and

$$\overline{Y} = A\overline{N}^{\alpha} \tag{B1.4.3}$$

Suppose that the wage level depends on the price level and the ratio of employment to the labor force:

$$W = \omega P^{\theta} P_{-1}^{1-\theta} \left(\frac{N}{\overline{N}}\right)^{\gamma} \text{ with } 0 \leq \theta \leq 1, \gamma \geq 0 \tag{B1.4.4}$$

In the short run, the wage level W is not fully indexed on the price level if $\theta < 1$. This is because wages are set by contracts that are renegotiated at discrete intervals. There is therefore *nominal wage rigidity* and a rise in the price level implies a drop in the real wage W/P. In the long run, however, there is full indexation as the wage adjusts to the price level and the real wage only depends on real factors. The wage level furthermore depends on the degree of tension on the labor market measured by $N/\overline{N}$, because a rise in employment improves the bargaining power of employees.

Supply is determined by the employers' profit-maximization behavior. The corresponding first-order condition is:

$$\frac{\partial Y}{\partial N} = \alpha A N^{\alpha-1} = \frac{W}{P} \tag{B1.4.5}$$

Combining (B1.4.4) with (B1.4.5) gives:

$$\alpha A N^{\alpha-1} = \omega \left(\frac{P_{-1}}{p}\right)^{1-\theta} \left(\frac{N}{\overline{N}}\right)^{\gamma} \tag{B1.4.6}$$

which gives a relation between employment and price, and therefore between production and price.

In the long run, $P = P_{-1}$ and the solution therefore implies $N = \overline{N}$, $Y = \overline{Y}$ and $\frac{W}{P} = \omega$. The supply curve is therefore vertical.

In the short run, however, P_{-1} is given and the solution is:

$$Y = HP^{\sigma} \tag{B1.4.7}$$

where H is a constant and

$$\sigma = \frac{\alpha(1-\theta)}{1+\gamma-\alpha} > 0$$

Production depends positively on price because a rise in the price level is only partially translated into a rise in the wage level and therefore reduces the real wage. The supply curve is thus upward-sloping and the price elasticity of supply depends negatively on the degree of short-term indexation of wage over price θ and on its responsiveness to employment level γ.

As evident in figure 1.6, the effectiveness of demand policies depends on the slope of the short-term supply curve. In an economy with widespread nominal rigidities and a low responsiveness of wages to labor market conditions, the short-run supply curve can be almost flat, which makes demand policies very effective. However, when indexation is fast and wages responsive to unemployment, the slope of the short-run supply curve can be almost vertical, rendering demand policies close to ineffective. So the choice of a policy response depends both on the identification of shocks and on the underlying properties of the economy.

This distinction is more delicate than it seems, however, because the structure of the economy is not known with certainty. In the less-than-perfect information world they live in, what economists do is to represent the structure of the economy by a model, in other words by a series of relationships between explained (left-hand-side) variables and explanatory (right-hand-side) variables, some of the latter being exogenous. To take a very simple representation, let a function F relate right-hand-side variables X to left-hand-side variables Y:

$$Y = F(X) \tag{1.3}$$

An observed change in the value of a Y variable can thus result from:

- A change in the value of the X variables, or
- A change in, or a perturbation to, the F relationship between X variables and Y variables

In real time, policymakers are seldom able to sort out with certainty the former from the latter. For example, they observe a rise in the price level but do not know whether it represents a normal response to shocks to input prices (e.g., oil) or results from an accelerated inflationary development.

A standard approach is to start from observation and estimate equations like $Y = F(X)$ with econometric techniques. For example, household consumption can be written:

$$C_t = a_0 + a_1 R_t + a_2 \frac{\Omega_t}{P_t} - a_3 \left(u_t - \bar{u}\right) + \varepsilon_t \tag{1.4}$$

where C is consumption, R real income, Ω nominal wealth, u unemployment, t designates time, and ε is the residual from the estimation (the error term that captures the difference between fitted and actual values of C). In principle, a change in C can result from:

- Changes in the values of the explanatory variables R, Ω, P, and u;
- A temporary shock to the equation, thus a change in ε, or
- A change in the a_i coefficients representing a durable modification of the structure of the economy.

Each of these three factors may call for a different policy response, if any.

Reconciling observation with our simple aggregate demand/aggregate supply framework raises further difficulties. First, the series of shocks ε_t depends on the estimated values of the a_i, in other words the identification of shocks is contingent on a particular representation of the economy. Second, the single-equation approach we have outlined allows separating out shocks, but if applied to GDP it fails to distinguish between supply and demand shocks, as both affect the residual. This is a problem, as the appropriate policy response depends on the identification of the shock. This requires more sophisticated techniques such as the one proposed by Olivier Blanchard and Danny Quah (1989), which builds on the fact that those shocks have opposite effects on quantity and price. They simultaneously estimate autoregressive equations linking endogenous variables such as output and price, and they treat the corresponding estimation residuals as exogenous shocks, which can be classified as demand or supply shocks. This, for example, allows determination of the origin of a slowdown in output.

Beyond these discussions, the effectiveness of macroeconomic policy has been the subject of an equally fierce controversy. Against the background of policy failures in the 1970s, economists and commentators have built on the advances of economic theory to claim that economic stabilization policy was inherently inefficient, despite the fact that such a result only holds under

specific assumptions (see chapter 2). Skepticism toward active stabilization policy remains widespread, especially in continental Europe.

c) Redistribution

As regards redistribution, the central argument for intervention is that even if the market-determined distribution of income is Pareto-optimal, this equilibrium does not necessarily ensure social justice. The prime motive for intervention here does not stem from a lack of efficiency of the market outcome—as for allocation and stabilization—but from a pure equity concern.

A normative criterion is generally required to decide what constitutes an improvement in equity. Which criteria can be used to compare two income distributions is the topic of the next section. What needs to be made clear immediately is that an "improvement" in equity—whatever is meant by that—can take place at constant efficiency, be traded off against a reduction in efficiency, or can trigger an increase in efficiency.

In the first case, equity concerns can be completely separated from efficiency ones. This happens when the government is able to modify the distribution of income through lump-sum transfers that do not affect economic incentives. Trade policy is a case in point: a classic result from trade theory is that under fairly general assumptions free trade (or more generally trade liberalization) improves overall efficiency and yields gains to all participating countries. However, the same trade theorems show that there are losers in the process: For example, labor loses and capital wins in a capital-rich country that opens to trade with capital-poor countries. Nevertheless, the overall gain from trade allows the government to redistribute the benefits from capital to labor in order to ensure that free trade is Pareto-superior to protection.

In practice, however, lump-sum transfers are almost impossible to implement. Take again the case of trade: to determine whom it should tax and to whom it should redistribute, the government would need to have full *ex ante* information on the effects of liberalization. Furthermore, it would need adequate instruments for redistributing. What it can do concretely is tax income, profit, or consumption and redistribute through targeted assistance programs or means-tested transfers. However, those taxes and transfers change economic incentives and therefore affect the market equilibrium. Equity cannot be separated from efficiency anymore.

This is why redistribution often involves an equity–efficiency trade-off: The more income is redistributed, the higher the efficiency loss, because both taxes and transfer reduce the supply of production factors (labor and capital). However, the opposite situation also exists and redistribution can in certain cases *improve* efficiency. For example, public policies aiming at ensuring access of the poor to education and health care frequently yield efficiency gains through improving the productivity of the labor force. Their justification thus goes beyond their equity effects.

1.3 Economic Policy Evaluation

1.3.1 Decision criteria

To evaluate economic policy choices, and especially to compare alternative policies, precise criteria are necessary. But can a single criterion be used for efficiency, stabilization, and equity? Although this is conceivable in theory, in practice economic policy choices are generally represented as implying trade-offs between different dimensions.

a) A single objective?

The most general purpose that can be assigned to economic policy is the satisfaction of resident households (in a political economy setup, one would say of voters), their *utility** as economists call it. In elementary textbooks, the consumer's utility depends on a limited range of items but nothing precludes broadening it. Determinants of household utility can obviously include the consumption of goods and services, the amount of leisure (and therefore, by difference, the quantity of labor supplied), and the quality of the environment. It is also possible to bring into play the variety of goods and services consumed, as well as altruistic or moral considerations (for example, the fact that a good was not produced using child labor).

For consumer i utility can be written, in a very general formulation:

$$U_t^i = U\left(C_{i1}^t, C_{i2}^t, \ldots C_{in}^t; N_i^t; E_i^t; \Xi^t\right) \tag{1.5}$$

where $C_{ik}^t (k = 1 \ldots n)$ is the amount of good k consumed by household i at time t, N_i^t the quantity of labor supplied by household i in period t, E_i^t a vector of variables representing working conditions (intensity of effort, painfulness ...) and Ξ^t a vector of variables representative of the quality of the environment.

Instantaneous utility is, however, insufficient. Based on such a criterion there would be no reason to invest (since investment increases the quantity of goods and services available for future consumption but reduces current consumption). Nor would there be reasons to prevent future global warming. An intertemporal approach is therefore needed. This requires defining a *discount rate** ρ[19] in order to aggregate utility over time:

$$U_i = \sum_{t=0}^{\infty} \frac{U_i^t}{(1+\rho)^t} \tag{1.6}$$

19. The discount rate ρ is the interest rate that should be paid to an agent holding a dollar for him to be indifferent between spending his dollar today and investing it at rate ρ. This is equivalent to saying that the agent is indifferent between receiving one dollar in a year and $1/(1+\rho)$ dollars today. $1/(1+\rho)$ is called the *discount factor**. On a perfect capital market, ρ is equal to the interest rate.

The intertemporal utility U_i of consumer i is thus the *present value** of her future utilities discounted at rate ρ. Although this representation remains very simple—for example, it completely overlooks uncertainty as regards the future or the possible irreversibility of some decisions—the simple fact that the sequence of all future utility levels can be taken into account greatly reduces the hedonistic character of the simple utility criterion. U_i indeed brings into play the future availability of goods and services. This criterion can be used to assess the desirability of structural reforms (box 1.5): it allows addressing the trade-off between present and future consumption or intertemporal trade-offs involving the preservation of natural resources whose availability will be valued by future generations. The same approach can be used for assessing the utility cost of policies that fail to keep the economy at long-term balance.

Much depends on the choice of the discount rate ρ: A high discount rate introduces a bias toward the short-term and immediate consumption; a low discount rate brings into play the welfare of future generations. This dimension is important as regards environment but also for economic policies having an impact on savings, such as tax and pension policies: as will be seen in chapter 6, saving is necessary for capital accumulation and therefore determines the long-term production level.

Box 1.5 Structural Reforms and Intertemporal Trade-offs

Structural reforms generally aim at medium-term effects. However, they also have a short-term impact. It can be positive (a tax reform often stimulates demand, especially if it involves tax cuts) or negative (the announcement of a future pension reform creates concern about the future, the reform itself leads households to re-examine their expenditure plans and can reduce consumption). Structural reforms therefore often involve intertemporal trade-offs.

The International Monetary Fund (2004) carried out an econometric study on the dynamic effects of structural reforms. It concluded that reforms of the labor market and to a lesser extent of the product market have negative short-term effects. Tax and financial reforms, on the other hand, have favorable short-term effects.

From a public economics standpoint, the decision criterion should be the present value of the net benefits from the reform. Thus, if V_t is the net increase in utility in period t of a reform carried out in period 0, a criterion for undertaking this reform is:

$$V = E\left(\sum_0^\infty \frac{V_t}{(1+\rho)^t}\right) \geq 0 \qquad \text{(B1.3.1)}$$

where E is the expectation operator and ρ the discount rate. V obviously depends on the discount rate chosen to compare benefits over time.

In public economics, it is the same as for any choice of investment. However, if the decision-maker has a strong preference for the short term, for example because he or she is subject to a re-election constraint, ρ is higher, which can result in discarding reforms that have positive medium-term effects but are expensive in the short term. Moreover, this evaluation is marred with uncertainty regarding future profits from the reform and their distribution over time. Risk aversion can also result in discarding reforms.

These problems lie at the core of the political economy of structural reforms. For example, trade liberalization brings medium-term efficiency gains (through a better resource allocation) but involves both short-term adjustment costs (because of implied industrial restructuring) and an immediate fall in tariff revenues.

This intertemporal utility function, however, remains that of a specific household or of a single, supposedly representative, household. The next step is to aggregate the utilities of heterogeneous individuals. This is fraught with difficulties: Must the utility of all agents be equally weighted? Can the well-being of some be reduced to increase that of others? Those questions have a long history in normative economics.

The *Pareto criterion**—according to which a policy improves upon the status quo if it increases the utility of at least one individual and does not reduce that of any other—only makes it possible to compare a limited set of situations and policies. Figure 1.7, borrowed from Atkinson and Stiglitz (1980), explains why. Let us consider two individuals 1 and 2, represent their respective utilities on the X and Y axes, and suppose that the AF locus gives all possible combinations of their respective utilities. According to the Pareto criterion, C is superior to any situation on AC and E is superior to any situation on EF, because moving to the North-East improves both utilities simultaneously. However, there is nothing we can say about the points located on EC.

Choice then requires a *social welfare function**:

$$\Gamma(U_1, U_2, \ldots, U_m) \tag{1.7}$$

where $1 \ldots m$ represent the individuals or households (or, more realistically, categories of households grouped, for example, by income deciles). This makes it possible to compare two income distributions and to decide which one is more desirable. The most usual functions are:

$$\text{the "Benthamian" function:} \quad \Gamma = U_1 + U_2 + \ldots + U_m,$$

and

$$\text{the "Rawlsian" function:} \quad \Gamma = \text{Min}(U_1, U_2, \ldots, U_m)$$

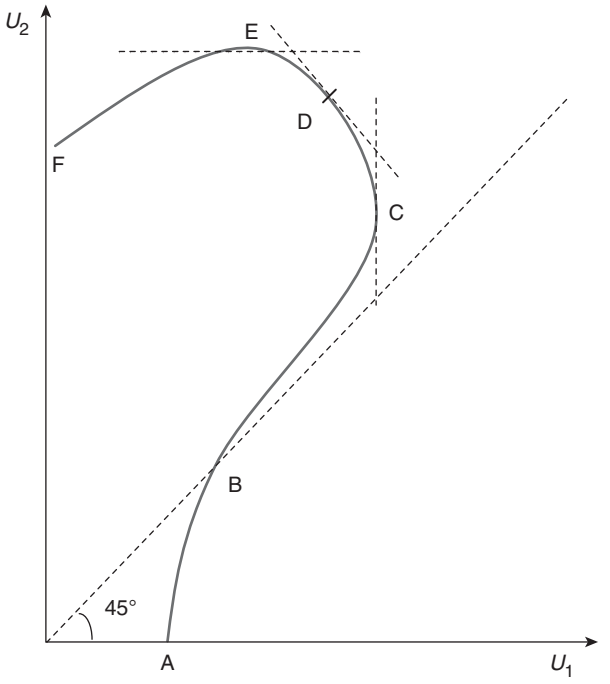


Figure 1.7 Individual utility and social choices: An illustration.
Source: Taken from Atkinson and Stiglitz (1980).

The first function is named after eighteenth-century philosopher and economist Jeremy Bentham.[20] It assumes that the distribution of utilities across individuals is of no importance and that only the aggregate utility matters. This results in choosing point *D* in figure 1.6 because it corresponds to the maximum aggregate utility (this is where the *AF* locus is tangent to a line of slope -1), even if the corresponding distribution of utility across individuals is definitely uneven as indicated by the distance to the bisector.[21]

Those who value social justice need a more equitable criterion. Strict equalitarianism would result in choosing *B* (intersection between the *AF* locus and the first bisector), which is not optimal according to the Pareto criterion. However, should simultaneous increases in the utility of both agents be rejected, only because they would not be equally distributed? A more satisfactory criterion, first proposed by John Rawls,[22] is to seek the

20. Jeremy Bentham (1748–1832) was the founder of utilitarianism.
21. It must, however, be pointed out that this social welfare function will likely exclude extreme income distributions, because the marginal utility of income decreases with income: a dollar given to the poor increases their utility more than it reduces that of the rich from whom it was taken.
22. The US philosopher John Rawls (1921–2002) authored *A Theory of Justice* (1971).

maximization of the utility of those who have the least of it. This *Maximin** principle leads to choosing C where utility U_1 of the least-favored individual is maximum.

It is therefore conceptually possible to assign to economic policy a single objective that includes the three motives for public intervention (allocation, stabilization, redistribution). However, this requires defining very general utility functions and deciding on their aggregation through time and across households. That would be supposed to have been decided *ex ante*, for all possible situations, on as complex an issue as the trade-off between intragenerational and intergenerational equity—how much the society stands ready to increase inequality at a certain point in time in exchange for an improvement in the welfare of future generations. It is unlikely that a society would be able to reach consensus on such a function.

b) Specific criteria for allocation, stabilization and redistribution

In practice, economic policy evaluation relies on separate, conceptually different instruments for allocation, stabilization, and redistribution assessments (box 1.6). Social welfare functions of the type presented above are generally used for evaluating allocation policies, but most often in a simplified form.

As regards allocation, *partial equilibrium** analyses are the simplest ones as they only consider the sector directly affected by policy decisions and neglect interdependence between sectors arising, on the demand side, from the economic agents' budget constraints, or on the supply side from the limited availability of production factors. For example, the analysis of the effects of reducing the indirect tax rate on a given good or service is limited to the market for that product and therefore overlooks the corresponding reduction of spending on other products and the overall impact of the implied reallocation of labor and capital that follows the increase in demand. Those are acceptable assumptions only to the extent that the sector considered is small in comparison to the whole economy.

Partial equilibrium analyses can be implemented easily as they only require knowing the price elasticity of supply and demand for the product considered and rely on well-known instruments such as consumer and producer *surplus** (an approximation of the variation of their utility). For example, a lowering of tariffs on imports generally reduces the surplus of local producers through increased competition from foreign producers, while consumers gain from lower prices (and the government loses from a reduction of tariff revenues). Standard economic theory predicts a net gain for the society that corresponds to the sum of the three agents' net surpluses. This calculation is valid, however, only if the sector is small. It is appropriate to use it for assessing the effects of eliminating a high tariff on a given product (a *tariff peak**, in the trade policy jargon), not for evaluating those of an across-the-board reduction on tariffs on manufactures.

When partial equilibrium analysis is inappropriate, evaluation must rely on a *general equilibrium** approach that takes into account interdependence across sectors and results in a situation where supply balances demand simultaneously on all markets. This can only be done with simulation models such as the *Computable General Equilibrium (CGE) models**, which are used for assessing the effects of complex trade, structural reform, or tax policy packages (box 1.6).

Box 1.6 Four Categories of Economic Policy Evaluation and Simulation Models

Computable General Equilibrium (CGE) Models

CGE models are based on an extensive representation of the economy with several categories of agents interacting on several markets. These models rely on an extensive description of economic decisions by households and firms that explicitly takes into account budget constraints and other accounting identities, frequently including input–output coefficients. Behavior results from explicit optimization. The corresponding parameters are not estimated from time series data (as in macroeconometric models) but they are *calibrated**—i.e., their values are chosen on the basis of *a priori* information and adjusted in order for the model to reproduce a given initial situation. This approach is preferred to econometric estimation, because the latter is generally impossible due to a very large number of parameters in comparison to available data.

The early CGE models of the 1960s and the 1970s were static and relied on simplifying assumptions as regards the structure of markets. Those currently in use are frequently dynamic and allow for imperfect competition and the absence of market clearing—for example, unemployment.

CGE models are widely used today whenever there is a need for evaluating the medium-term consequences of policy decisions affecting several markets or several agents simultaneously. They are the standard instrument for evaluating the outcome of trade negotiations conducted multilaterally within the framework of the WTO or at the regional level (see, for example, GTAP—*Global Trade Analysis Project*—on www.gtap.agecon.purdue.edu/). They are also the premier instrument for assessing the economic effect of environmental policies—for example, those aiming at reducing greenhouse-gas emissions. Other fields of application include development economics (especially for countries where time series on past behavior are unreliable or irrelevant because

they have undergone major reforms) and economic history (where CGE techniques can be used to assess the effect of events and decisions). The strength of CGE models rests on their comprehensiveness, their internal consistency, and the fact that they are based on explicit optimizing behavior. They can also be highly disaggregated and therefore take into account differences across sectors or categories of households. Their weakness is that they do not adequately represent short-run effects and rely on weak empirical bases.[23] Major international institutions such as the OECD or the World Bank have developed CGE models or rely on those developed by university research.

Macroeconomic Models

Macroeconomic models were initially derived from Keynesian theory, of which they were intended to provide a formal and quantified representation. They have gradually evolved in line with developments in macroeconomic theory and empirical techniques. They are used both for forecasting and policy simulation purposes.

The main variables in a macroeconomic model (e.g., consumption, investment, employment, price-setting, foreign trade) are taken from national accounts; corresponding behavior is determined by structural equations. The equation parameters are generally *estimated** with econometric techniques[24] or alternatively calibrated.

The approach originates in the first models built by Jan Tinbergen and Lawrence Klein in the 1950s. Initial macroeconomic modeling was essentially empiricist but it gradually introduced more theoretical discipline, partly in response to a flurry of academic critiques (see chapter 2). In order to respond to the criticism that they were assuming very naïve behavior on the part of private agents, model-builders introduced model-consistent expectations about the future values of model variables, thereby renouncing the initial (implicit) assumption that they had better knowledge of economic behavior than the agents themselves (see table B1.6.1). In response to the criticism that their models were based on ad-hoc assumptions and lacked theoretical underpinnings, they increasingly developed explicit microeconomic foundations for the

23. See Shoven and Walley (1984) for a survey of CGE modeling. A recent example is given by Lofgren et al. (2001).

24. In its simplest form, econometric estimation consists in determining the parameters of an equation linking a dependent variable to observed explanatory variables in such a way that the deviation of estimated from actual values of the dependent variable is minimum. A popular technique is the Ordinary Least Squares (OLS) estimation, which consists in finding the parameter values that minimize the sum of square deviations of estimated from actual values of the dependent variable over the estimation period.

estimated equations and adopted more rigorous estimation techniques. At the same time, multinational modeling was developed in order to provide a representation of international interdependence.

Estimated macroeconomic models provide readily available instruments for assessing the impacts of shocks or policy decisions and they are therefore still widely used, despite having been subjected to scathing critiques. When used with care, they provide useful coarse-cut estimates of policy effects. They are widely used in government administrations, central banks (including the US Federal Reserve, the Bank of Japan, and the European Central Bank), international institutions (OECD, IMF, European Commission), and forecasting institutes (NIESR in the UK, etc).

Table B1.6.1
Four generations of macroeconomic models

Model type	Keynesian adaptive expectations	Keynesian rational expectations	Real business cycle	Dynamic stochastic general equilibrium
Strengths	Allows assessment of the impact of policies and shocks in a unified manner	Generates more realistic dynamic responses to cyclical disturbances	Strong theoretical foundations, improved supply side	Integrates aggregate supply and demand responses through microeconomic theory
Weaknesses	Adaptive expectations allowed policymakers to consistently mislead others, creating a bias toward expansionary policies	Absence of strong theoretical foundations made it difficult to assess effects of policies on aggregate supply	Assumption of flexible prices left little room for analysis of macroeconomic policies	Models are in early stages of development and large ones are difficult to build and run

Source: Adapted from International Monetary Fund (2004).

A new generation of macroeconomic models called *Dynamic Stochastic General Equilibrium* (*DSGE*)* models has been developed in the 1990s and 2000s in response to dissatisfaction with both the short-run limitations of the general equilibrium models and the long-run properties of the macroeconomic models. This new approach builds on the insights of the real business cycle models of the 1980s but explicitly introduces

nominal rigidities in the Keynesian tradition. Thus, consumers maximize intertemporal utility and producers maximize intertemporal profit, but sticky prices prevent markets from clearing.

DSGE models bridge the gap between CGE and macroeconomic models. They include both "deep parameters" (akin to those of general equilibrium models—which are either calibrated or estimated while taking into account *a priori* information on their expected values) and standard estimated parameters.

DSGE models were initially developed in academic research but have recently been adopted by institutions such as the International Monetary Fund (Botman et al., 2007) and the European Central Bank (Smets and Wouters, 2003), where DSGE modeling supplements traditional macroeconometric modeling.

Statistical Models

Statistical models depart from the *a priori* hypotheses about agents' behavior that characterize CGE and macroeconometric models. These models were first developed in the 1980s in response to dissatisfaction with macroeconometric models (see notably Sims, 1980). Their aim is to empirically determine interdependences between endogenous variables by estimating simultaneously several equations without a priori theoretical restrictions. *Vector Auto Regressive models* or *VARs** are specified in autoregressive form, which implies that each variable depends on its own past values as well as on those of other variables. For example, the effects of monetary policy are assessed through estimating simultaneously the dependence of GDP, inflation, and the short-term interest rate on their past values. Some parameter restrictions derived from theory can be introduced in so-called structural VARs but they are kept at a minimum.

VARs and structural VARs are frequently used for assessing the effects of macroeconomic shocks and policy changes, such as exchange-rate shocks and monetary policy decisions, and they tend to substitute larger-scale macroeconometric models for such purposes (see chapter 4). However, their very aggregate character does not allow them to be used for more detailed policy analyses.

Other examples of statistical models are *factor models**, where the joint dynamics of a large set of short-term economic indicators (such as industrial output, prices, household and company survey data, etc.), which are typically observed at a monthly frequency, is assumed to derive from a smaller number of underlying, hidden variables called factors. Sargent and Sims (1977) find that two dynamic factors explain more than 80% of the variance of a number of economic variables, including the rate of unemployment, wholesale price inflation, growth in industrial production, employment growth, etc. These models are used by central

banks and economic institutes to produce forecasts and better anticipate turning points in economic sentiment. The current US Federal Reserve chairman, Ben Bernanke, has himself contributed to developing this approach (see, for example, Bernanke and Boivin, 2003), and is a strong advocate of the development and use of a dynamic factor model within the Federal Reserve to improve its forecasts.

Microsimulation Models

Even detailed CGE models make simplifying assumptions as regards the categories of agents represented in the model. Yet for the assessment of tax or social policy measures, what is needed is an evaluation that takes into account heterogeneity among households. This is what microsimulation models aim at through explicitly representing a large number of categories of households or individuals.

Those models build on the development of large-scale databases providing information on individual agents and can include individual information on tens of thousands of persons, if not more. Equations typically combine optimization (as regards, for example, labor supply decisions), calibration (as regards, for example, the evolution of an individual's employment status resulting from the probability of losing one's job or of finding a new one when unemployed), and econometric estimation (as regards, for example, estimated wage equations determining an individual's wage as resulting from her or his age, gender, and human capital).

Microsimulation models have the great advantage of providing information that allows assessment of the distributional effects of policy changes. However, they do not provide an evaluation of their macroeconomic effects. These models are widely used for assessing the impact of changes in tax and welfare benefit legislation. Examples include the European EUROMOD model based at the University of Essex or the TAXBEN model of the London-based Institute for Fiscal Studies, a simplified version of which is available on the web.

Social welfare functions could also be used to evaluate the effectiveness of stabilization policies. This relies on the assumption that a single agent's utility suffices to represent the social cost of a departure from equilibrium. Also, trade-offs between short-term stabilization and long-term allocation can be evaluated provided the social welfare function has an intertemporal dimension. A major difficulty, however, arises from measuring the welfare loss from unemployment: in a microeconomic setting, voluntary unemployment increases individual utility because agents value leisure, yet it is difficult to claim that a rise in unemployment increases utility. Another difficulty comes

from the cost of inflation. In a microeconomic setting, expected inflation is neutral and it does not affect utility, provided agents do not hold significant nominal balances.[25] Yet it seems absurd to argue that a combination of unemployment and inflation has no effect on welfare or even increases welfare. Therefore, the analysis of stabilization policies generally relies on specific *macroeconomic loss functions* such as:

$$L_t = E_t\left(\sum_{s=0}^{\infty}(1+\rho)^{-s}\sum_{i=1}^{N}\alpha_i(y_{t+s}^i - \tilde{y}^i)^2\right) \tag{1.8}$$

where $E_t(X)$ stands for the mathematical expectation at date t of variable X; the y_i are the objectives of economic policy (typically, growth and inflation) and $\tilde{y}_i$ are the corresponding target values (which can in theory be derived from optimization behavior); α_i is the weight assigned to variable i; and ρ is a *discount factor**. The objective of the government or the monetary authorities is to minimize the value of the loss function.

In practice, policymakers never use such functions (most finance ministers would be surprised to see them), but this representation is a fair approximation of reality. Decision-making processes do bring trade-offs into play: For example, between reducing the budget deficit and bolstering GDP growth or between supporting consumption and promoting investment. The inflation–growth trade-off was a key concern in the 1960s and the 1970s, and the desire to avoid being confronted with it again exerted a considerable influence in the choice of an institutional architecture that assigns responsibility for monetary policy to an independent central bank (see chapter 4). This type of reasoning, moreover, is encouraged by the recourse to models for decision-making. In a way, the representation of economic policy choices in simulation instruments retroacts on economic policy.

The analysis of stabilization policies generally consists in comparing, with the same loss function, policy reactions to a given *shock**—an exogenous event such as a fall of world growth or a variation of the investors' appetite for risk. The loss function allows one, for example, to determine whether, in response to an adverse shock to private investment, it is preferable to increase public investment, reduce corporate taxation, or lower the interest rate. Results, of course, depend on the macroeconomic model and on the loss function used.

As regards redistribution, social welfare functions are almost never used to support concrete decisions. Discussions on the redistribution effects of economic policies are almost always based on empirical indicators of

25. In a rational expectations, neoclassical framework, the welfare costs of inflation only arise from the costs of individuals holding cash (inflation implies a penalty to holding cash, which necessitates that agents go more often to the bank to withhold cash, in turn requiring larger banks and a greater number of person-hours, etc.) and the costs of changing price tags (i.e., circulating changing information about prices). These two series of costs are respectively called shoe-leather costs (as an illustration of the premature wear and tear of shoe soles as consumers need to go more often to the bank) and menu costs. See Pakko's (1998) review on shoe-leather costs.

inequalities, such as the distribution of income between *deciles** of population, or aggregate indicators such as the *Lorenz curve** and *Gini coefficient** (box 1.7). This is because income levels are more palatable and natural references in policy discussions and public debates than utility—although it must also be recognized that relying exclusively on the comparison of income levels can be misleading (for example, for a given distribution of income an increase of subsidies to social housing is likely to improve the utility of individuals in the bottom deciles while public investment in higher education increases the utility of those in the top deciles).

Box 1.7 Measuring Inequality

The simplest and most telling measure of income (or wealth) inequality is the ratio of the income (or wealth) of the top 10% of the population to the bottom 10%. According to the Human Development Report (United Nations Development Program, 2005), it stands at 6.2 in Sweden, 15.9 in the US, 94 in Brazil, and reaches 103 for the world as a whole. It is this type of measure that is frequently used in public debate.[26] However, to summarize the whole distribution by the gap between the two extremes overlooks developments affecting 80% of the population.

The *Lorenz curve* provides a graphic representation of the entire distribution. Fractiles of the population ordered by income level are plotted on the X axis and the corresponding cumulative share of total income on the Y axis. For an (x, y) point on the curve, y is therefore the share of total income going to the first x% of the population. The bisector corresponds to an equal distribution of income. The greater the distance between the Lorenz curve and the bisector, the larger the inequalities. Figure B1.7.1 gives Lorenz curves for US pre-tax family income in 1980, 1990, and 2005. It is apparent that income inequality has widened.

The *Gini coefficient* provides a synthetic numerical measure of inequality. It is defined as twice the surface of the area between the Lorenz curve and the bisector, which is comprised between 0 (perfectly equal distribution) and 0.5 (maximum inequality). The Gini coefficient therefore varies between zero and 1. Formally, if $x_i (i = 1, \ldots n)$ are the limits of the fractiles of the population and y_i the share of each fractile in total income, the Gini coefficient is:

$$G = 1 - \sum_{i=1}^{n} \left(x_i - x_{i-1}\right)\left(y_{i-1} + y_i\right)$$

26. Research pioneered by Thomas Piketty has relied on a similar approach to analyze the evolution of the share of very high incomes (the top one percent or the top one per thousand) in national income. It provides evidence of a significant rise in the share of top incomes since the 1980s in the US and the UK, while the same phenomenon has not been observed in continental Europe. See Atkinson and Piketty (2007) and for the US case Piketty and Saez (2003).

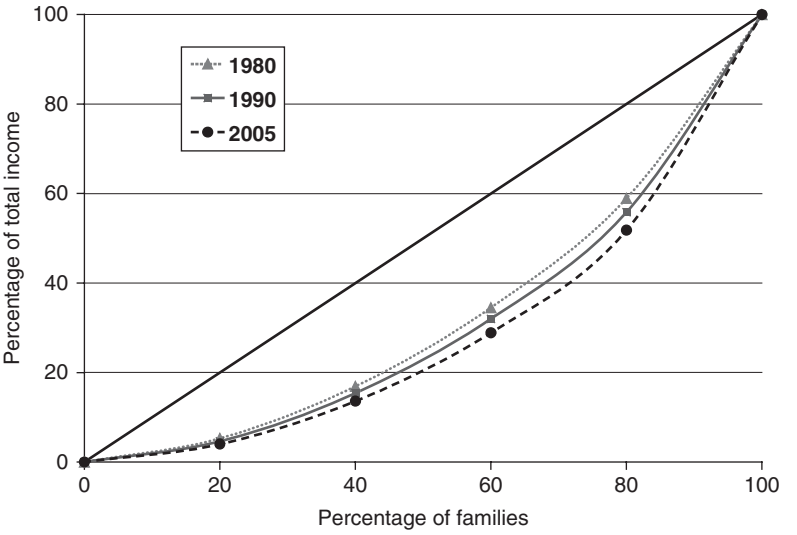


Figure B1.7.1 Lorenz curve, US families, 1980–2005.
Source: US Census Bureau.

In the late 1990s and the 2000s, the Scandinavian countries and Japan were those where Gini coefficients were the lowest (figure B1.7.2). They were highest in South America and some African countries.

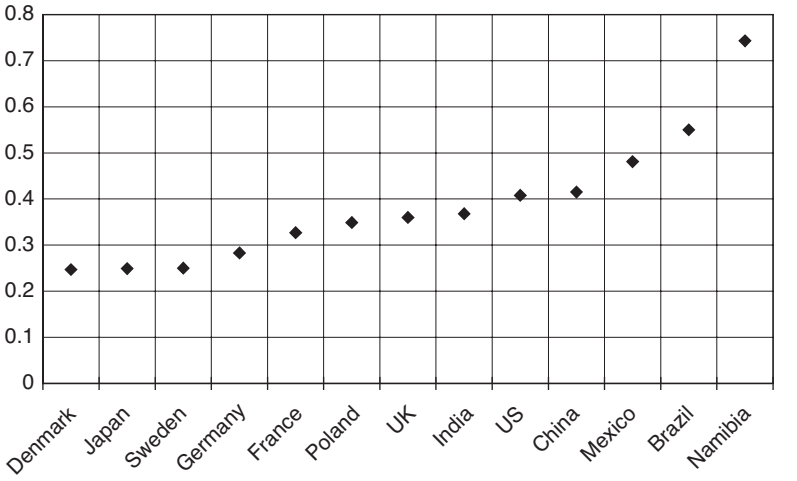


Figure B1.7.2 Inequality among individuals: Gini coefficients, selected countries and groupings, 2000s.
Source: UNDP Report, 2005.

1.3.2 *Ex post* evaluation and experiments

Whatever the criterion used, model-based policy evaluation is of an *ex ante* nature, as it typically compares the current situation to the one that is expected to result from a given policy. Even when implemented *ex post* to compare the situation after the implementation of a given policy to the one that would have prevailed had this policy not been implemented, the evaluation is based on model parameters previously estimated or calibrated. Therefore, it does not take into account information from changes brought by the policy under scrutiny. In fact, there is no difference between an *ex ante* and an *ex post* model-based policy evaluation—but for the presentation of the results.

a) Making use of experiments

Ex ante evaluation is especially inappropriate when the reform has a structural character and is expected to modify behavior in a way that does not simply replicate past experience (this remark, which economists call the Lucas critique, will be developed later on). It is therefore important to carry out genuine *ex post* policy evaluations. Building on a standard practice in life sciences, corresponding techniques have developed in the 1990s, especially in the field of social policies. They often rest on *natural experiments** that make it possible to compare the behavior of individuals affected by the policy change to that of other individuals whose situation, though similar, has not been affected by it. This is, for example, the case for means-tested transfers: By comparing the behavior of individuals immediately below the threshold to that of those immediately above it—which are therefore very similar in all relevant dimensions—it is possible to measure precisely the effect of the policy (box 1.8). In certain countries, policymakers also resort to *controlled experiments** to assess the potential effect of a policy change under consideration. For example, Canada used this technique to evaluate the employment effect of in-work benefits before they were introduced.[27] MIT economist Esther Duflo received the 2010 Clark medal for her work on Randomized Control Trials (RCT). In RCTs, the policy subject to evaluation is assigned randomly so that there is no selection bias in the constitution of treatment and control groups.

Natural and controlled experiments are used in a variety of fields, from taxation and social transfers to education and punishment of criminals. Though experiments are standard practice in research, their utilization by policymakers remains uneven.

27. Concretely, a sample of potential beneficiaries was randomly selected and divided into two groups. The first one was offered the new in-work benefits, while the second one served as control group. Comparison between the employment behaviour of the two groups served to determine the effect of the scheme. See Michalopoulos (2002).

Box 1.8 Evaluating Public Policies through Natural Experiments

The traditional method for evaluating the response of labor supply to tax changes is to resort to econometric estimates on time series. It is fraught with methodological difficulties, from the identification of labor supply to the lack of experience with relevant tax changes. Also, particular tax changes may affect only certain categories of the labor force and aggregate estimates do not allow assessment of the corresponding effects.

The issue would be easy to solve were it possible to resort to controlled experiments—as in the life sciences—with laboratory techniques. This would involve selecting a group of individuals, submitting them to a tax change, and observing their behavior in comparison to a pilot group with comparable characteristics for whom the taxation would have been left unchanged. Such an experiment would make it possible to isolate the pure effect of taxation.

Experiments of this kind are practiced in certain countries, such as the US, Canada, or The Netherlands. They are used to evaluate the effectiveness of envisaged social policy reforms before they are generalized. However, in other countries, such as France, the practice of controlled experiments was for long held back by constitutional difficulties.

An alternative is to exploit *natural experiments*, as, for example, when two jurisdictions within the same country which previously had similar legislation start implementing differing policies. This closely replicates the conditions of a controlled experiment; comparison of the resulting behavior allows evaluation of the effectiveness of the different policies. This holds even when the two jurisdictions did not have the same policies: The effect of introducing a new policy can be assessed by comparing changes after it has been introduced (this is called the *difference in differences method**). Even within centralized states where legislation is uniform, some events can be regarded as natural experiments. For example, Joshua Angrist and Victor Lavy (1999) were able to make use of the rule that in Israeli public schools a class must be divided into two when its size reaches 40. This rule generates exogenous variations in class size which can be used to study the effect of class size on the pupils' performance.

The econometric techniques in use for analyzing natural experiments were first developed by James Heckman (2000). They aim at eliminating the effect of heterogeneities and selection biases between the target and the control populations. The diffusion of these methods in the 1990s has led to a major advance in the evaluation of social policies.

b) Evaluation criteria in practice

In practice, policy evaluation frequently relies on crude criteria for measuring the effect of a decision on, for example, the *Gross Domestic Product** (GDP) (i.e., of the total value added to products in the economy during a year),

on unemployment, or on various income groups. Some of these criteria lack rigorous economic foundations. This especially applies to GDP: An increase in defense expenditure or in spending on security devices in response to a terrorist threat may increase GDP but does not increase welfare (in comparison to the situation that prevailed prior to the threat). In a full-employment situation, a reduction in working time (an increase in leisure), may increase welfare, but reduces GDP. Even a decline in unemployment does not necessarily improve welfare if, for example, it is obtained at the price of a reduction of the job search period and leads to a deterioration of the matching between labor supply and labor demand. To have more people at work but more of them unhappy and less productive than they could have been, had they spent a few more weeks looking for a suitable job, can hardly be regarded as an improvement. Alternative criteria have been developed to better measure well-being and happiness (see box 1.9). In September, 2009, an International Commission chaired by Nobel Prizewinner Joseph Stiglitz (Stiglitz et al., 2009) documented the many defaults of GDP and made recommendations to develop indicators that better account for welfare heterogeneity across individuals and for sustainability.

Box 1.9 Economic Development and Human Development

The economist and philosopher Amartya Sen (1999) has pointed out that the life expectancy of African-Americans is lower than that of inhabitants of the Indian state of Kerala. This illustrates how money income can be a misleading indicator of living conditions. In reaction to the deficiencies of GDP per person, new indices have been developed such as the Human Development Index and other composite indices by the Human Development Report office of the United Nations Development Program. Those indicators take into account a number of health, education, and social criteria (nutrition, life expectancy, access to health care, etc.). Although initially rather crude, this approach has gradually gained in sophistication, in large part thanks to Sen's research. In the late 1990s, it inspired the definition and adoption by the international community of the *Millenium Development Goals*, which set a number of concrete and measurable social objectives for 2015. The Human Development Index (HDI) introduced in 1990 is a composite index whose calculation involves life expectancy at birth, knowledge (as measured by adult literacy rate with a two-thirds weight, and the combined primary, secondary, and tertiary gross enrollment ratio with one-third weight) and GDP per capita in Purchasing Power Parity[28] US$ (for a detailed explanation, see Technical note 1 in the United Nations Development Program (2006) Human Development Report).

28. As market exchange rates exhibit wide fluctuations, statisticians often use Purchasing Power Parity (PPP) exchange rates for international comparison purposes, PPP exchange rates are

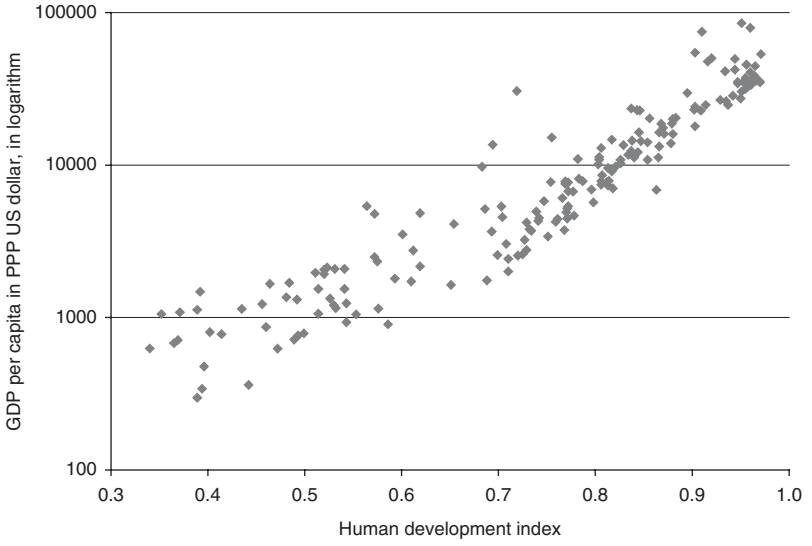


Figure B1.9.1 GDP per capita and human development index in 182 countries, 2007.
Source: UNDP, Human Development Report 2009.

As shown in figure B1.9.1, the aggregate Human Development Index is strongly correlated to the economic development level measured by GDP per capita. Nevertheless, significant exceptions do exist: producers of energy and raw materials, such as the Gulf States, exhibit a lower HDI ranking than their GDP ranking, which suggests high inequality and insufficient provision of public goods such as health and education; in contrast, some poor countries reach relatively high levels of human development. For example, the HDI of the seven countries in the 1900 PPP$ per capita to 2100 PPP$ per capita income bracket in 2007 ranged from 0.377 (Angola) to 0.688 (Vietnam).

One difficulty with the HDI is how to aggregate different items such as GDP per capita and life expectancy. An alternative approach to measuring economic development consists in starting from an indicator of income and in adjusting it through several corrections, including health, education, leisure, risk of becoming unemployed, etc. For example, Fleurbaey and Gaulier (2007) evaluate citizens' implicit willingness to

computed on the basis of international price level surveys. They are intended to ensure that a unit of currency A converted at PPP exchange rate into currency B retains the same purchasing power, i.e., can buy the same basket of goods and services. If P^A and P^B are the corresponding price levels, the PPP exchange rate Q is such that $(QP/P^*) = 1$. For further explanations and discussion, see chapter 5.

pay for a number of improvements in the quality of life (as compared to a common standard) and then correct the gross national income per capita for the corresponding amounts. They thus produce some sort of a living-standard-income-equivalent that allows comparison of countries. This kind of approach can only be applied to countries of similar standard of living because it relies on marginal effects. Table B1.9.1 shows how such corrections affect the ranking of OECD countries in terms of standard of living.

Table B1.9.1
Rankings according to GDP per head and income adjusted for living standards (2004, in Purchasing Power Parity US dollars)

	GDP/capita		Adjusted income
Luxembourg	68719	Luxembourg	55828
Ireland	40058	Norway	39975
United States	39518	Ireland	39782
Norway	38288	Japan	34989
Switzerland	33541	Austria	34695
Iceland	33090	Switzerland	33701
Austria	32176	US	33315
The Netherlands	32056	France	32805
Denmark	31974	Iceland	31972
Canada	31129	The Netherlands	31348
Belgium	31009	Italy	30442
United Kingdom	30843	Denmark	29689
Australia	30116	UKm	29233
Finland	29816	Canada	28414
Japan	29539	Belgium	28366
Sweden	29499	Spain	28131
France	29077	Sweden	28027
Italy	28162	Germany	27276
Germany	28147	Australia	26508
Spain	25341	Finland	26034
New Zealand	22912	Greece	22582
Greece	21954	Korea	21653
Korea	20371	New Zealand	21320
Portugal	19687	Portugal	19163

Source: Fleurbaey and Gaulier (2009).

1.3.3 Collateral effects

So far, we have treated the allocation, stabilization, and redistribution functions separately. In reality, an economic policy decision often has effects in more than one dimension. For example, a cut in personal income tax

has a permanent allocation effect (it increases labor supply), a temporary stabilization effect (it increases the private agents' income and therefore their demand for goods), and a sustained redistribution effect (it improves the relative income of the agents in the highest income brackets).

It often happens that a policy is adopted for its positive effects on one dimension even though it has adverse effects on other ones, giving rise to trade-offs:

- Redistribution policies frequently introduce undesirable distortions in resource allocation. Means-tested social transfers (such as minimum income) serve a distributive objective but often create *inactivity traps**,[29] and therefore reduce labor supply;
- Trade opening is generally pursued for its allocative effects (the gains from specialization and the corresponding productivity effects, technology spillovers associated with foreign direct investment, etc.) but also has effects on the distribution of income as unskilled jobs are relocated as a consequence of trade with developing countries. The same is true of technical change;
- A reduction in inflation (stabilization policy) can have undesired effects as regards the distribution of income (redistribution) because economic agents are unequally able to protect their income during disinflation. It can also affect allocation if the unemployment resulting from anti-inflation tightening becomes persistent.

However, a policy adopted for one motive can also have positive effects on other dimensions. For example, a redistribution policy aiming to improve the net pay of unskilled workers (through a workers' tax credit or cuts in social contributions) can have favorable allocation effects through a rise in labor supply.

Finally, the sign of the effect is not always clear. The link between inequalities and growth provides an example. Income inequality is sometimes claimed to be positively correlated with growth, either because it allows part of the population to save and accumulate capital, or because innovation creates rents which benefit the innovators. The evolution of inequality within China illustrates this relationship. However, inequality is also claimed to be harmful to growth because it does not allow the poorest segments of the population to have access to education and health and it increases the risk of social and political disruption. The standard example here is Latin America.

Table 1.1 summarizes some of these interdependences.

29. An inactivity trap arises when the recipient of a state-dependent or means-tested replacement income (unemployment allowance, welfare transfers) has weak or non-existent economic incentive to return to work because the loss of social benefits makes the monetary gain from taking up a job too low to compensate for the reduction in leisure.

Table 1.1
Direct and indirect effects of three public policies (direct effects are indicated in **bold** type)

	Allocation	Stabilization	Redistribution
Reduction in income tax	+ **(increase in labor supply)**	+ (increase in demand for goods)	− (increase in inequalities)
Increase in government expenditures	+/− (depends on the content of expenditure and on the possibility of crowding out private expenditure)	+ **(by hypothesis)**	+/− (depends on the content of expenditure)
Increase in social transfers	− (risk of inactivity trap)	+ (increase in the demand for goods)	+ **(reduction in inequalities)**

Note: The initial situation is supposed to be characterized by Keynesian unemployment.

Conclusion

We have outlined in this chapter what economic policy aims at and which instruments it relies on. However, we have not explained why it is a matter for disagreements. The evidence, however, is that economic policy controversies abound. As encapsulated by the motto of Bill Clinton's 1992 presidential campaign ("it's the economy, stupid"), and again by Barack Obama's campaign in 2008, a large part of electoral campaigns are generally fought on economic matters. So why is it that reasonable people may disagree on economic policy?

This chapter provides some answers or at least some hints. Politicians can first pursue different social welfare functions: they may, for example, hold contrasting views about the desired distribution of income. Second, they can respond differently when confronted with trade-offs, for instance between equality and efficiency. Third, they may discount differently tomorrow's welfare, that is, they may have different time preferences. Those three dimensions of genuine policy preferences, attitudes toward trade-offs, and time preferences go a long way toward explaining familiar disputes between left-wing and right-wing parties.

The same type of reasoning provides clues as to why supposedly neutral bodies such as the international institutions are confronted by often-strident opposition from nongovernmental organizations. Ironically indeed, the term "Washington consensus" was coined in 1989 by John Williamson to designate

a set of policy prescriptions that "more or less everyone in Washington would agree were needed more or less everywhere in Latin America".[30] In the event, it soon came to designate a set of prescriptions a large part of opinion strongly disagreed with.

Ravi Kanbur (2001), a development economist who worked at the World Bank, has tried to shed light on the nature of disagreements on international economic policy choices. He posits that they can arise from differences in the *level of aggregation* adopted, the *time horizon* considered, and assumptions made on *market structure and power*. This especially applies to the debate between proponents and opponents of globalization:

- *Aggregation*: proponents emphasize the aggregate welfare gains from trade openness, because income redistribution can be corrected by fiscal transfers. However, opponents doubt that such corrective policies will actually be implemented and they fear that the benefits of globalization will accrue to the few and not to the many.
- *Time horizon*: proponents have a medium-term horizon of five-to-ten years and they neglect both the very short term and the very long term; opponents insist on short-term adjustment costs (in particular for the poorest, which relates to the previous point) and on long-term sustainability.
- *Market structure*: proponents generally suppose that markets are competitive and cleared by prices; opponents underline their imperfection and point out that market openness without government intervention has an adverse impact on income.

Kanbur's third item introduces a dimension that has not been addressed in this chapter but will be taken up in chapter 2, namely the uncertainty about the structure and functioning of the economy and the resulting policy disagreements. Although advances in economic knowledge have gradually reduced the scope for traditional disputes, new controversies have appeared. For instance, the growth and employment effects of tax policy are a matter for disagreement. Such controversies abound and regularly impact on the policy debate, although frequently in a distorted way.

Furthermore, there are additional reasons for disagreements that go beyond either the choice of policy objectives or the uncertainty about instrument efficiency. To understand why, we will need to depart from the somewhat simplistic vision of what economic policy is about that provided the intellectual framework of most of this chapter. This is also taken up in chapter 2.

In concluding, however, it is worth recalling that politicians remain free to ignore what economists think is true. Most economists would, for example, say that protracted budget deficits eventually raise long-term interest rates because they increase the supply of Treasury bonds, but then US Vice-President Dick

30. See John Williamson's 2004 account of the history of the Washington consensus.

Cheney reportedly cut short discussions by saying that "Deficits don't matter. Reagan proved that".

References

Akerlof, G.A. (1970), "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics*, 84, pp. 488–500.

Angrist, J., and V. Lavy (1999), "Using Maimonides' Rule to Estimate the Effect of Class Size on Student Achievement," *The Quarterly Journal of Economics*, 114, pp. 533–74.

Arrow, K., and G. Debreu (1954), "Existence of an Equilibrium for a Competitive Economy," *Econometrica*, 22, pp. 265–290.

Atkinson, A., and T. Piketty (2007), *Top Incomes over the Twentieth Century*, Oxford University Press.

Atkinson, A., and J. Stiglitz (1980), *Lectures on Public Economics*, McGraw Hill.

Bernanke, B.S., and J. Boivin (2003), "Monetary Policy in a Data-Rich Environment," *Journal of Monetary Economics*, 50, pp. 525–46.

Blanchard, O., and D. Quah (1989), "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review*, 79, pp. 655–73.

Blanchard, O., and J. Tirole (2008), "The Joint Design of Unemployment Insurance and Employment Protection: A First Pass," *Journal of the European Economic Association* 6(1), pp. 45–77.

Botman, D., Ph. Karam, D. Laxton, and D. Rose (2007), "DSGE Modeling at the Fund: Applications and Further Developments," *IMF Working Paper no. 07/200*, August.

Brainard, W. (1967), "Uncertainty and the Effectiveness of Policy," *American Economic Review*, 57 (May, Papers and Proceedings), pp. 411–25.

Buchanan J. (1975), "A Contractarian Paradigm for Applying Economic Policy," *American Economic Review*, 65, pp. 225–30.

Chapman, B. (2006), "International Reforms in Higher Education Financing: Income Related Loans," in Hanushek, E., and F. Welch, eds., *Handbook of the Economics of Higher Education*, Elsevier/North-Holland.

Dixit, A. (1996), *The Making of Economic Policy*, MIT Press.

Fleurbaey, M., and G. Gaulier (2007), "International Comparisons of Living Standards by Equivalent Incomes," *CEPII working paper 2007–03*, January.

Galí, J., and P. Rabanal (2004), "Technology Shocks and Aggregate Fluctuations: How Well Does the RBS Model Fit Postwar US Data?," *NBER Working Papers 10636*, National Bureau of Economic Research.

Heckman, J. (2000), "Microdata, Heterogeneity, and the Evaluation of Public Policy," Nobel conference [www.nobel.se].

International Monetary Fund (2004), "GEM: A New International Macroeconomic Model," mimeo, www.imf.org.

Kanbur, R. (2001), "Economic Policy, Distribution and Poverty: The Nature of Disagreements," *World Development*, 29, pp. 1083–94.

Kemp, M., and H. Wan (1976), "An Elementary Proposition Concerning the Formation of Customs Unions," *Journal of International Economics*, 6, pp. 95–97.

Keynes, J.M. (1931), *Essays in Persuasion*, Harcourt, Brace and Company [www.uqac.ca/].

Keynes, J.M. (1936), *The General Theory of Employment, Interest and Money*, Macmillan Cambridge University Press.
Kindelberger, C. (1978), *Manias, Panics and Crashes: A History of Financial Crises*, Wiley.
Klemperer, P. (2004), *Auctions: Theory and Practice*, Princeton University Press.
Kydland, F.E., and E.C. Prescott (1982), "Time to Build and Aggregate Fluctuations," *Econometrica*, 50(6), pp. 1345–70.
Lofgren, H., R. Louis Harris, and S. Robinson (2001), "A Standard Computable General Equilibrium (CGE) Model in GAMS," *International Food Policy Research Institute (IFPRI) Discussion Paper no. 75*, available at www.ifpri.org.
Michalopoulos, Ch., D. Tattrie, C. Miller, P.K. Robins, P. Morris, D. Gyarmati, C. Redcross, K. Foley, and R. Ford, (2002), "Making Work Pay: Final Report on the Self-Sufficiency Project for Long-Term Welfare Recipients," http://www.mdrc.org/publications/46/abstract.html.
Minsky, H. (1992), "The Financial Instability Hypothesis," Jerome Levy Economics Institute Working Paper No. 74, May.
Musgrave, R., and P. Musgrave (1989), *Public Finance in Theory and Practice*, McGraw Hill.
North, D. (1993), "Economic Performance Through Time," Nobel Prize lecture, 9 December, http://www.nobelprize.org.
Pakko, M.R. (1998), "Shoe-Leather Costs of Inflation and Policy Credibility," *Federal Reserve Bank of St Louis Review*, November/December.
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), pp. 283–99.
Piketty, T., and E. Saez (2003), "Income Inequality in the United States, 1913–1998," *The Quarterly Journal of Economics*, 118(1), pp. 1–39, updated tables available on http://elsa.berkeley.edu/~saez/.
Rawls, J. (1971), *A Theory of Justice,* The Belknap Press of Harvard University Press.
Rousseau, J.J. (1755), "Economie," in Diderot, D., and J. le Rond d'Alembert, *Encyclopédie*, vol. 5, pp. 337–49. English translation: http://quod.lib.umich.edu/d/did/, Scholarly Publishing Office of the University of Michigan Library, 2009.
Sargent, T.J., and C.A. Sims (1977), "Business cycle modelling without pretending to have too much a-priori economic theory," in Sims, C., et al. eds., *New Methods in Business Cycle Research*, Federal Reserve Bank of Minneapolis.
Sen, A. (1999), *Development as Freedom*, Oxford University Press.
Shiller, R.J. (2003), *The New Financial Order: Risk in the 21st Century*, Princeton University Press.
Shoven, J.B., and J. Whalley (1984), "Applied General Equilibrium Models of Taxation and International Trade: An Introduction and Survey," *Journal of Economic Literature*, 22, pp. 1007–51.
Sims, C. (1980), "Macroeconomics and Reality," *Econometrica*, XLVIII, pp. 1–48.
Smets, F., and R. Wouters (2003), "An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area," *Journal of the European Economic Association* 1, pp. 1123–75.
Stein, H. (1986), *Washington Bedtime Stories*, Free Press.
Stiglitz, J., and A. Weiss (1981), "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, 71(3), pp. 393–410.

Stiglitz, J., A. Sen, and J.P. Fitoussi (2009), Report of the Commission on the Measurement of Economic Performance and Social Progress, http://www.ofce.sciences-po.fr/pdf/documents/rapport.pdf.

Tinbergen, J. (1952), *On the Theory of Economic Policy*, North-Holland.

United Nations Development Program (2005), *Human Development Report*, United Nations.

United Nations Development Program (2006), *Human Development Report*, United Nations.

Viner, J. (1950), *The Customs Union Issue*, New York, Carnegie Endowment for International Peace.

Williamson, J. (2004), "A Short History of the Washington Consensus," paper prepared for the Fundación CIDOB conference "From the Washington Consensus toward a new Global Governance," Barcelona, Spain, September 24–25.

2

Economic Policy in a Complex World

2.1 Living with limits
- 2.1.1 The limits of knowledge
- 2.1.2 The limits of representation
- 2.1.3 The limits of confidence
- 2.1.4 The limits of information
- 2.1.5 The limits of benevolence
- 2.1.6 The policy responses

2.2 Living with interdependence
- 2.2.1 The rise of interdependence
- 2.2.2 International policy coordination
- 2.2.3 Federalism

Conclusion

References

In chapter 1, economic policy was presented as an engineer's science. A single omniscient, omnipotent and benevolent policymaker was supposed to engage in optimization, taking social preferences as given, and relying for decisions on accurately estimated parameters. It is now time to challenge those assumptions.

We begin by questioning a series of assumptions that are implicit in this representation. Since the 1970s, economic research has systematically explored the deficiencies of the traditional approach to economic policymaking, pointing out severe theoretical and empirical limitations and exploring what remains of the standard prescriptions when those limitations are lifted. As expressed in very similar terms by Avinash Dixit (1996) and Jean-Jacques Laffont (2000c), the research program of the last three decades can be read in retrospect as knocking the omniscient, omnipotent, and benevolent policymaker's statue off its pedestal.

What is important, however, is not only to realize the extent of the criticism. As argued in this chapter and in the following ones, it is also important to understand how to make the most of economic policy in

a complex and imperfect world. The theories developed since the 1970s model the decision-maker as interacting in an imperfect information context with other players, who are themselves imperfectly informed but are able to anticipate, compute, and play, and whose behavior depends on their expectation of the decision-maker's not-always-benevolent actions. Economic policy can still influence the players, but they can no longer be regarded as preprogrammed automatons. This representation has deeply influenced the theory and, gradually, the practice of economic policy. Contrary to what superficial analysis would suggest, the impact of these developments has not primarily been a downgrading of economic policy; rather, it has changed its design, and the governance technologies it relies on.

The second main issue we address in this chapter is interdependence, both between countries and between sub-levels of government. Over the last few decades, economic policy has been deeply affected by the rise of international interdependence. Most notably in Europe, but also elsewhere, the ever more frequent assignment of policy responsibilities to higher (supra-national) as well as lower (regional or local) levels of government has made the model of the central government as a single policy-player increasingly irrelevant. Economic policy must today be regarded as a multiplayer game structured by the vertical relationship between levels of government. This is certainly more true in Europe than in the US, which was a federal country from the start and does not easily accept encroachments on national sovereignty. However, even in the US, the trend is discernable.

This chapter is intended to provide a basis for the policy-specific chapters that follow. Section 2.1 surveys various limitations of the traditional description of economic policy, and outlines their consequences for the design and implementation of government intervention. Section 2.2 discusses the making of economic policy when various levels of government interact, with a special focus on the European Union and on global governance.

2.1 Living with Limits

There are five main limits to the traditional approach to economic policy. First, governments have imperfect knowledge of the structure of the economy and of future risks. Second, firms and households are not akin to ants under a magnifying glass: They devise their own strategies and they react to—and anticipate—economic policy measures. Third, policymakers may not be able to convince private agents that they will actually do what they have announced, and this affects the behavior of private agents. Fourth, policymakers may not have the information they need to take decisions. Fifth and finally, policymakers may not pursue the general interest. In what follows, we look at each of those limits in turn, before discussing how economic policy has developed tools to address them.

2.1.1 The limits of knowledge

An important—though implicit—assumption in most of chapter 1 was that the government has extensive knowledge of the preferences of economic agents and of the structure of the economy. That public and private agents invest in the acquisition of information and make use of the information they have is certainly a natural assumption, but it has limits. We will explore here four, nonmutually exclusive positions:

- The parameters of models used by economists and decision-makers are fraught with uncertainty.
- Decision-makers usually base their decisions on expected outcomes only and seldom take into account the full distribution of risks.
- Rare but very damaging events are a challenge for policy decisions, but the distribution of risks is usually not well known, and in some cases it cannot even be quantified using traditional probabilistic methods.
- In an uncertain environment, there are situations in which it is preferable to wait before acting: There is a "precautionary principle" of economic policymaking (this same principle can also, however, justify prompt action instead of waiting in some circumstances).

a) Model and parameter uncertainty

Let us start from a simple representation of the economy:

$$\boldsymbol{Y}_t = H(\boldsymbol{X}_t, \boldsymbol{Y}_{t-1}, \boldsymbol{Y}_{t-2} \ldots, \boldsymbol{\theta}, \boldsymbol{\varepsilon}_t) \tag{2.1}$$

where $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$ are multidimensional vectors respectively summarizing government action, policy objectives, parameters, and random shocks that are out of the reach of the government. $\boldsymbol{X}_t$ can for example represent tax rates and public expenditures at time t, while $\boldsymbol{Y}_t$ represents household consumption, and θ the elasticity of consumption with respect to income, wealth, and the interest rate. H represents the accounting and behavioral relations linking all variables, and $\boldsymbol{\varepsilon}_t$ is a random vector, the value of which is unknown until period t. Barring unexpected shocks, the state of the economy at time t thus depends on its past evolution and on current government actions. However, their impact depends on $\boldsymbol{\theta}$ which is not directly observable. There are two sources of uncertainty about $\boldsymbol{\theta}$:

- First, *model uncertainty** arising from the choices made by the theorist and the econometrician. Questions here are, for example: Should the interest rate be included in the consumption function? Or, are consumption, investment, and export functions linear? There are many choices that model builders can make, given the theoretical assumptions. Policymakers are not always aware that the analyses and recommendations they are presented with rest heavily on model choices by econometricians.

- Second, for a given model, *parameter uncertainty**, arising from the limited range of observed data available to the econometrician. What is available for policy analysis is not the true value of $\boldsymbol{\theta}$ but an estimate $\hat{\boldsymbol{\theta}}$ extracted from individual or time-series observations with the help of more–or-less-sophisticated econometric techniques (box 2.1). Policymakers are usually not aware of the extent to which $\hat{\boldsymbol{\theta}}$ is fraught with uncertainty. For instance, a government facing a recession will feel comfortable undertaking output stimulation through increased public spending if the Keynesian multiplier (i.e., the reaction of output to a given increase in public spending) is known to be higher than one. Econometric evidence does suggest that point estimates of the multiplier are close to one but, given the distribution of the estimate, there are substantial odds that it could be lower than zero (see chapter 3).

Box 2.1 Parameter Uncertainty in Econometric Models

Suppose that the economy is governed by the following equation (the dependence on Y_{t-1} has been dropped for the sake of simplicity):

$$Y_t = H(X_t, \theta, \varepsilon_t) \tag{B2.1.1}$$

If all relationships are linear, this can be rewritten as:

$$Y_t = \theta X_t + \varepsilon_t \tag{B2.1.2}$$

If Y comprises n variables and there are m exogenous variables in X, θ is an $m \times n$ matrix of parameters to be estimated based on the observed values of X and Y over the period $t = 1$ to T. The *ordinary least squares* (OLS)* estimate of θ, that is, the value of θ that minimizes the sum of squared residuals, is:

$$\hat{\theta} = (\Xi'\Xi)^{-1}\Xi'\Psi \tag{B2.1.3}$$

Where Ψ is a $T \times n$ matrix built by staggering the observed values of Y_t and Ξ is a $T \times m$ matrix built by staggering the observed values X_t. Since X and Y are random variables, $\hat{\theta}$ is also random and converges only asymptotically (i.e., when there are a very large number of observations) toward the true value θ. The same is true when parameters are observed not only in the time dimension but also across individual observations.

The variance–covariance matrix of $\hat{\theta}$ can be computed as a function of Ψ and Ξ. For example, if Y includes only one variable (for example GDP), if the m variables in X are deterministic and if the variance of the random shock ε is constant over time and equal to σ^2, the variance of $\hat{\theta}$ is a $m \times m$ matrix:

$$Var\ \hat{\theta} = \sigma^2(\Xi'\Xi)^{-1} \tag{B2.1.4}$$

More generally, one can compute the variance of any well-behaved function of $\hat{\theta}$:

$$Var\, g(\hat{\theta}) = \sigma^2 \frac{\partial g}{\partial \theta} (\Xi' \Xi)^{-1} \frac{\partial g}{\partial \theta}, \tag{B2.15}$$

This allows building *confidence intervals** for $g(\hat{\theta})$ (i.e., ranges of values for a given confidence level). For instance, suppose that the point estimate of the Keynesian multiplier is equal to unity, with a 90% confidence band of [0; 2] and a 95% confidence band of [−0.5; 2.5]. This would mean that there is a 90% probability that the multiplier takes a value between 0 and 2, and a 95% probability that it ranges from −0.5 to 2.5. The larger the band for a given level of confidence (say 95%), the lower the reliability of the point estimate 1. In such an example, it would not be possible to say at a 95% confidence level whether the multiplier is positive or negative.

b) Risk

In most instances, private companies do a better job than the public sector of taking into account the distribution of risks in their decisions. The head of marketing who launches a new product and the credit officer who extends a credit to a company do not make their decisions on the basis of expected profit only. They appreciate the possibility that the project might fail or that the credit would not be refunded, so they provision for this risk or require the appropriate collateral. A telling and widely used measure of possible loss is *Cost at Risk** (CaR) which measures how much may be lost at a given confidence level. As an example, if the unit return of an investment project is random and follows a normal law with mean 1 and standard deviation 2, then the expected return of investing €1 million is €1 million, the loss in 10% of cases is more than €1.55 million, and the loss in 30% of cases is more than €50000. This results from the cumulated distribution of returns, shown in figure 2.1.

The same method is used in capital markets to assess the maximum loss of value out of a financial asset or of an asset portfolio at a given time horizon, in which case it is called *Value at Risk* (VaR)*. VaR is the cornerstone of modern risk management in financial institutions and requires knowledge of the joint distribution of the returns of all underlying assets.

Since von Neumann and Morgenstern (1944) who formalized the seminal work of Daniel Bernoulli (1738), economists have generally assumed that agents know the probability of the various states of nature and maximize the expected value of their future utility, i.e., the average of utility in each state of nature weighted by its probability. Within this framework, the instrument used to model attitudes toward risk is *risk aversion**, which is related to the second derivative of the utility function (box 2.2).

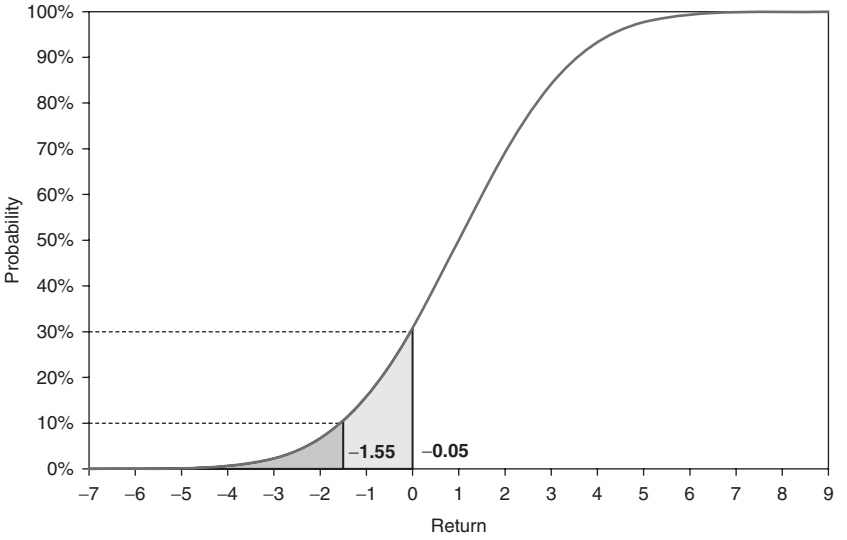


Figure 2.1 Cost at risk.
Reading: There is a 30% probability that the return is below −0.05.

Box 2.2 Risk Aversion

Suppose that a representative household utility increases with income, but that the marginal utility of income is a decreasing function of income. This is a standard assumption in consumption theory, which is supported by empirical studies on the relationship between income and happiness (Layard, 2005). In mathematical terms this corresponds to $U'(R) > 0$ and $U''(R) < 0$ where R is income and U is utility.

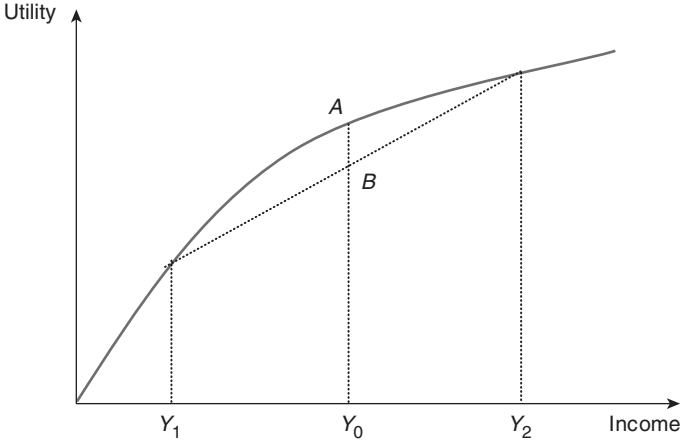


Figure B2.2.1 Utility and income.

Assume that the household is given the choice between receiving income Y_0, and receiving with a probability of 0.5 either Y_1 or Y_2 such that $Y_0 = (Y_1 + Y_2)/2$. Clearly, expected income is the same in both cases, but since utility is a concave function of income, utility is higher in the first case (corresponding to point A on the graph) than in the latter case (point B). The household prefers certainty to uncertainty. This is risk aversion.

Intuitively, risk aversion depends on the concavity of the curve depicting the relationship between income and utility, i.e., on $U''(R)$. However, this is not a workable definition since utility is invariant with respect to a strictly monotonic transformation.

Two common definitions of risk aversion are therefore used. The first is *absolute risk aversion** (ARA) and the second *relative risk aversion** (RRA) also called the *Arrow–Pratt coefficient**. They are defined by:

$$ARA = -\frac{U''(R)}{U'(R)} \tag{B2.2.1}$$

and

$$RRA = -\frac{RU''(R)}{U'(R)} \tag{B2.2.2}$$

It is usual to use functional forms for utility that exhibit constant absolute risk aversion (CARA) or constant relative risk aversion (CRRA). For example, with logarithmic utility ($U(R) = \log R$) the relative risk aversion is constant and equals one. A more general form is:

$$U(R) = \frac{R^{1-\alpha} - 1}{\alpha - 1} \tag{B2.2.3}$$

with $\alpha \neq 1$.

In this case, the relative risk aversion is constant and equals α. Risk aversion is a basic concept in the theory of consumption and savings and has many applications in finance.

In many economic models, companies are assumed to be risk-neutral (because they have access to financial markets) and individuals are assumed to be risk-averse. It is assumed that companies maximize the net present value of their expected future profits, while individuals maximize the present value of the *expected utility* of their future consumption. It is assumed that utility is a concave function of consumption, which implies that individuals' welfare decreases when expected future consumption remains constant but its uncertainty increases. When the utility function exhibits constant relative-risk aversion, it can be shown that a higher variance of shocks to future

consumption induces *precautionary savings** (see Deaton, 1992, for a review of consumption).[1]

In recent time, crises like hurricane Katrina have forced governments to take risk into account; but when they focus on risk, policymakers sometimes ignore expected outcomes. Think of the following example in the field of public health (Gollier, 2001): In 2000, the French government had to choose between two new compulsory tests, of which one tested for HIV and hepatitis C in batches of transfusion blood, and the other for breast cancer among over-50-year-old women. Independent research pointed to a cost per saved year of life of about €9 million for the former and €1500 for the latter. However, because of public sensitivity to transfusion-related diseases, the French government wanted to show it was doing its best to prevent blood contamination and chose the former test in spite of the available cost–benefit analysis.

c) Extreme or unquantifiable risks

This leads us to another issue, which has to do with the distribution of risk. Most economic models rest on the assumption that shocks are normally distributed, i.e., that their distribution has the well-known "bell curve" shape, with a given mean and standard error. There are, however, circumstances in which this assumption cannot hold: Shocks may be skewed, in which case their median value is not equal to their mean, or their distribution may exhibit *fat tails**, meaning that very rare events are more likely to occur than under a normal distribution.[2] For instance, it has usually been supposed since Louis Bachelier's seminal work on the French bond market that financial asset returns follow a normal probability distribution (Bachelier, 1900/2006). For financial economists this is a very convenient assumption, but, in practice, it is not valid and the 2007–09 financial meltdown provided a powerful example of an extreme financial risk. As once noticed by Benoît Mandelbrot, there have been 48 days in the period 1916–2003 when the Dow Jones Industrial Average, a US stock index, moved by more than 7% in a single day, an event which should occur once every 300,000 years in a normal distribution (Mandelbrot and Hudson, 2004). Mandelbrot has advocated using a more general class

1. Recent research in experimental economics has however challenged the expected utility paradigm. Experiments show that individuals do not adhere to rational decision-making behavior, that they frequently rely on rules of thumb rather than complex evaluations, that they try to avoid losses—including sometimes through taking more risk—and that their choices depend on initial conditions and the framing of decisions. Research in this field is very active but its results have not yet been incorporated in economic models.

2. The empirical measure of the "fat-tailedness" of a variable X is its *kurtosis** defined as $k = \mu_4/\sigma^4$ where $\mu_4 = E(X - EX)^4$ is the fourth order moment of the variable and $\sigma^4 = [E(X - EX)^2]^2$ is the square of its variance. For a normal distribution, it can be shown that $k = 3$. A distribution is said to be *leptokurtic**, or fat-tailed, if $k > 3$.

of distributions, the *Pareto–Levy distributions**, which exhibit fat tails and sometimes do not even have a finite variance.

Rare events are all the more important when there are nonlinear mechanisms at play in the economy. For instance, a particularly pronounced recession can throw the economy into a state of deflation (chapter 4) in which traditional monetary policy instruments will be inefficient and will increase the duration of unemployment up to a point where many laid-off workers will lose their skills and will never be hired again. Conversely, a very high inflation rate will initiate wage-indexation schemes that will be very difficult to repeal. The policy conclusion is that central banks can tolerate reasonable fluctuations of inflation around the target value but they must be quite vigilant to avoid extreme risks (Svensson, 2004). In other words, they should pay special attention to events whose probability is low but whose disruptive effects are high, or *tail risks**.

Diseases such as HIV/AIDS, BSE (Bovine Spongiform Encephalopathy, the "mad cow" disease), or avian flu are examples of extreme risks that make it very difficult for governments to devise and calibrate a policy response. In a period of globalization and rapid technical progress, there are many such examples. To guide decision-making, it is crucial to obtain a consensus on the probabilities of various risks, based on independent expertise.[3]

It is difficult to assess the probability of extreme events such as wars, natural disasters, or change of political regime. For example, no available model in physical oceanography makes it possible to quantify the risk that global warming might lead to an inversion of the Gulf Stream, an event which would have far-reaching consequences on both sides of the Atlantic. In 1921, Frank Knight distinguished between *risk**, when randomness can be described by a probability measure, and *uncertainty**, when it cannot. Under uncertainty, traditional economic models are useless since they rely on expected utility *à la* Von Neumann and Morgenstern (1944). Knightian uncertainty has been applied to financial asset pricing (Epstein and Wang, 1995) and to game theory, but it is not widely used for policy analysis purposes, and most economists still do not make the distinction between "risk" and "uncertainty".

d) The option value of waiting

A last criticism that can be made of the traditional approach to economic policy in an uncertain environment is that it focuses in great detail on the substance of policy decisions, while the major question is often that of their timing. The key concept here is *irreversibility**. If all policy decisions were incremental and reversible, economic policy would be state-contingent: It would adapt at any point in time to the current state of the economy. However, in a world where decisions have irreversible consequences, it can be optimal to wait until new information is available on their cost and benefits.

3. On the economics of extreme events, see Posner (2004).

A well-known result from the theory of investment under uncertainty is that, since the decision to invest is irreversible while the decision to defer investment is reversible, proper economic calculation implies comparing the value of investing today to the value of investing at any other possible point in the future (Arrow, 1968, McDonald and Siegel, 1986). This implies that, as a rule, investment should be undertaken only if benefits exceed costs by a certain amount, which is an increasing function of the variance of the project's return. In other words the possibility of deferring the project has a value, analogous to that of a financial option. One speaks of an *option value** attached to the project.

This concept has a very wide scope and applies to all decisions that involve irreversibility, fixed costs, or discrete (as opposed to continuous) choice in an uncertain environment. Infrastructure investments are a straightforward example, but the same reasoning can also be used at a macroeconomic level. In 1997, when they had to decide on adopting the euro, a typically irreversible decision, the UK and Sweden chose to exercise the option to "wait and see" how the eurozone would perform (see also box 5.13 on the five tests set by the UK government).

As noticed early on by Claude Henry (1974), this approach is particularly relevant in assessing projects that cause irreversible damage to the environment, such as building roads through forests, burying nuclear waste below ground, or drilling for oil in a nature reserve. The choice is even more complex when doing nothing may also involve irreversible consequences, as in the case of climate change. Limiting carbon dioxide emissions requires massive investment with an uncertain return and a known opportunity cost, which suggests a policy of waiting. However, annual emissions continue to increase the concentration of carbon dioxide in the Earth's atmosphere (carbon dioxide decays extremely slowly), and there is wide suspicion among scientists that a persistently high concentration may cause large-scale, nonlinear events such as bifurcations in climate dynamics. Inaction thus increases the cost of future stabilization, and doing nothing has a cost of its own. Policymakers have to make a trade-off between investment irreversibility and environmental irreversibility.[4]

Facing these dilemmas, one would like to delineate a *precautionary principle** for economic policy, as for environmental problems.[5] There are several pitfalls, however. First, it may not be possible to obtain consensus on a common metrics to gauge costs and benefits, such as increased consumption on the one hand and environmental or human damage on the other. Second, choosing the rate at which future costs and benefits should be discounted

4. On the application of option theory to climate change, see Ha-Duong (1998).

5. The precautionary principle was introduced at the United Nations Rio Conference on Environment and Development. Article 15 of the Rio Declaration states that "where there are threats of serious and irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation." The precautionary principle was enshrined in the treaty of Lisbon of the European Union.

raises difficult theoretical issues with a very long time horizon where no market interest rate is available. The general theoretical answer is that long-time-horizon discount rates depend on the shape of the utility function.[6]

How to tackle this problem is at the core of controversies about climate change policies. The Stern Report commissioned by the British government (Stern, 2007) argues that in spite of the costs involved in policies that aim at containing climate change there should be no waiting, because not taking immediate action could involve high economic costs in the future. It estimates that the present value of the welfare costs involved in mitigating greenhouse gas emissions equals those of a permanent reduction in the level of world GDP by one percentage point, whereas the present value of the welfare costs of inaction would equal that of a permanent reduction of GDP by five to 20 percentage points. On the other hand, its critics claim that Lord Stern does not properly discount future costs (Nordhaus, 2007; Weitzman, 2007), and that he underestimates the potential for reducing the cost of mitigation policies through technical progress and innovation. This first argument raises difficult issues in comparing welfare across generations (box 2.3). The latter argument builds on the discussion on the option value of waiting.

Box 2.3 Climate Change and the Discount Rate

The evaluation of policy options for climate change mitigation involves assessing costs and benefits over very long time horizons, typically 50 years or more. The result of such assessments heavily depends on the discount rate used for the computation of present values. For example, the present value of a dollar paid in 50 years time is 60 cents with a 1% discount rate, but it is only 14 cents with a 4% discount rate. This implies that in the first case an optimizing policymaker would be ready to pay 60 cents upfront to prevent one dollar of damage in 50 years, but that it would only be ready to spend 14 cents in the second case.

How to discount the future in presence of uncertainty and the possibility of major damage is a challenge to standard intertemporal optimization. As developed by Guesnerie (2003), the intuition that each generation has a responsibility toward future ones can be justified formally in a model where private consumption of standard goods and the environmental good are only partially substitutable, and leads to choosing a near-zero time discount rate. Consistent with this approach, the Stern report takes, on philosophical grounds, the view that the welfare of all

6. More specifically, the case for a precautionary principle depends on the relationship between *prudence**, defined as the third derivative of the utility function and absolute risk aversion (Gollier et al., 2000; Gollier, 2002).

generations should have the same weight because there is no reason to value our offspring's welfare less than our own. When computing the present value of the welfare of all future generations', it uses a near-zero time discount rate δ.[7] The welfare function to maximize is thus:

$$W = \sum_{t=0}^{\infty} \frac{U(C_t)}{(1+\delta)^t} \tag{B2.3.1}$$

where $t = 0, 1, 2 \ldots$ represent generations, C_t is the consumption of generation t, and $U(C_t)$ the corresponding utility. Equation (B2.3.1) can be rewritten in continuous time, which is mathematically more tractable:

$$W = \int_0^{\infty} U(C_t) e^{-\delta t} dt \tag{B2.3.2}$$

Even with $\delta = 0$, however, intertemporal welfare maximization does not imply that all generations' *consumption* should be valued equally. This is because in the presence of technical progress, future generations will have access to higher levels of consumption. As the marginal utility of income and consumption is decreasing (see box 2.2), it remains desirable to trade a lower consumption in the future against a higher consumption today. In concrete terms, the reduction in consumption made necessary by the mitigation of damage to the climate should take into account that development is expected to make future generations better off.

Intertemporal maximization therefore leads to discounting future consumption at a rate that is normally higher than the pure time discount rate δ. Formally r, the rate at which future *consumption* should be discounted, equals δ, the rate at which future *utility* should be discounted, plus the product of the rate of technical progress g, by the elasticity of the marginal utility of consumption with respect to consumption α.

$$r = \delta + g\alpha \tag{B2.3.3}$$

This is known as the *Ramsey equation** (after Frank Ramsey, an early twentieth-century economist who established the mathematical conditions for optimal growth, see chapter 6). The distinction between the *pure time discount rate** δ and the *social discount rate** (of consumption) r is an important one.

Stern takes $\delta = 0.1\%$, $g = 1.3\%$, and $\alpha = 1$ (which corresponds to a logarithmic utility function), which yields $r = 1.4\%$. This is indeed very low in comparison to discount rates commonly in use.

This choice of parameters has given rise to a controversy. Nordhaus (2007) claims that Stern's approach is disputable and that very different

7. Technically, Stern uses $\delta = 0.1\%$ to take into account the probability of extinction of mankind.

results would have been obtained with a discount rate derived from market interest rates or with other intertemporal social welfare functions, such as a Rawlsian one that would maximize the welfare of the poorest generation, or one that would aim at maximizing the minimum consumption along the riskiest time path (see chapter 1).

In response Stern (2008) criticizes the "inappropriate application of a marginal method in a strongly nonmarginal context" and points out that the solution suggested by his critics—implicitly to invest the money and spend it on solving the climate problem later—ignores irreversibilities and the fact that in a multi-good context, the price of environmental goods will likely have gone up sharply enough to make the standard one-good calculation inadequate.

e) Implications for policy

Uncertainty and risk have strong potential policy implications. Many errors have been made because governments based policy on wrong parameter estimates or did not properly take risk and uncertainty into account. Policy thinking is increasingly attentive to these issues.

An elementary example, which is telling because it illustrates how easily policy can be wrong even with the simplest dimensions of uncertainty, is provided by fiscal policies. Member states in the EU are obliged to release yearly medium-term budgetary plans called *Stability Programmes*. The European Commission (2007) has used this information to describe *ex post* deviations of the fiscal deficit from targets set three years earlier. The results of this evaluation are given in table 2.1 for the 15 countries that were members of the EU over the 1998–2006 period. While wisdom would have called for basing budgetary plans on prudent forecasts, it is apparent that this was not done. In spite of higher-than-expected revenues, frequent expenditure overruns have resulted in the deficit exceeding target in two-thirds of the cases. It is thus apparent that, on average, European governments do not manage public finances in a prudent way to take into account risks to revenues or expenditures. Some do: The Dutch government, for example, deliberately uses underestimated GDP forecasts to build its fiscal plans, which is a very rough way to take uncertainty on board.

Central banks deal with uncertainty too—and they are increasingly describing their role in terms of a decision-under-uncertainty framework. In the US, the Federal Reserve has become increasingly aware of the existence of tail risks and has altered its policy stance correspondingly. According to Frederic Mishkin, a Fed governor, the usual representation of policy based on a linear model and a quadratic loss function (the representation that was introduced in chapter 1, see section 1.3.1) "may provide a reasonable approximation to how monetary policy should operate under fairly normal

Table 2.1
Surprises and outturns in 15 EU budgetary plans, 1998–2006

	Nominal GDP growth, %	Nominal government revenue growth, growth, %	Nominal government expenditure growth, %	Nominal government balance as percentage of GDP, %
"Positive" surprises	50	58	76	36
"Negative" surprises	50	42	24	64

Note: The table gives for each variable the frequency of "positive" surprises (higher-than-forecast result) or "negative" surprises (lower-than-forecast result). The sample consists of all programs submitted by the 15 EU member states over the 1998–2006 period.

Source: European Commission (2007).

circumstances," but in the presence of tail risks, "optimal monetary policy may also be nonlinear and will tend to focus on risk management" (Mishkin, 2008). Aggressive easing by the Federal Reserve in early 2008 illustrates this philosophy which Fed Chairman Ben Bernanke (2008) summarized by saying that the Fed would "act in a timely manner as needed to support growth and to provide adequate insurance against downside risks." The reference to the *insurance* function of monetary policy is by no means inadvertent. The importance of tail risk has been increasingly recognized by central banks and international institutions such as the International Monetary Fund (Lipsky, 2008).

2.1.2 The limits of representation

In the previous section, we have highlighted the existence of uncertainty about the value of parameters, and have concluded that this should lead policymakers to exercise caution before they take decisions. However, we did not question the policymakers' ability to obtain unbiased estimates of the parameters. Public intervention becomes even more questionable if based on systematically inaccurate parameter values.

While Keynesianism had reigned supreme throughout the 1960s, the last three decades of the twentieth century were marked by a heated debate on the rationale, the methods, and the limitations of public intervention. The discussion was ignited in the early 1970s by a series of sharp criticisms of the goals and methods of economic policy. These criticisms came primarily from economists who objected to the very principle of government intervention and, vindicated by the failure of macroeconomic policies to achieve their primary goals of output stabilization and price stability, especially after the first oil shock, had embarked on a far-reaching project to debunk the inconsistencies of traditional approaches.

Building on the development of macroeconomic modeling in the 1960s, a simple and telling image of economic policy dominated in the 1970s. The modeling of human behavior had made it possible to develop a seemingly reliable representation of how a given policy variable would impact economic variables based on equation (2.1). Within this framework, economic policy consisted in selecting the value of X that minimizes (generally under some constraints) a loss function $L(Y)$ ascribing relative weights to the various policy objectives. The respective roles of the policymaker and the economist were then clear: The former's responsibility was to choose L, the latter's role was to identify H and estimate $\boldsymbol{\theta}$—the optimal economic policy then followed (see chapter 1).

a) Rational expectations

This paradigm was first challenged by John Muth. In a technical article published in 1961 in *Econometrica*, he introduced the notion of *rational expectations**. In models used in that time, the expectations of households and company managers regarding the future values of economic variables were often disregarded. When they were taken into account, they were assumed to be extrapolated from the last observed trends. Expected future inflation, which matters for consumption, saving, or wage negotiation, was for example supposed to depend on the observation of inflation over the past months or years. Muth showed that this assumption amounts to supposing that agents do not use all information available to them at the time of the decision, and are therefore not rational. Rational agents would instead make use of all available information, including about current and expected policy action, and forecasting errors would result only from events that were not foreseeable (box 2.4).

Box 2.4 Modeling Expectations

The expectation $Z^a_{t,t+1}$ made at time t of the value of variable Z at time $t+1$ can be written as a function of its present and past values as well as of other relevant variables $\boldsymbol{X}$:

$$Z^a_{t,t+1} = G\left(Z_t, Z_{t-1}, Z_{t-2}, \ldots X_t, X_{t-1}, X_{t-2} \ldots\right) \qquad \text{(B2.4.1)}$$

This formulation covers a number of possible specifications. One can suppose that individuals expect economic variables to revert to some long-term equilibrium value, or on the contrary to amplify past movements.

A convenient specification is *adaptive expectations**, which incorporates new information gradually:

$$Z^a_{t,t+1} = (1-\lambda)Z^a_{t-1,t} + \lambda Z_t, \quad 0 < \lambda < 1 \qquad \text{(B2.4.2)}$$

If $\lambda = 1$, then an adaptive expectation simplifies into a *static expectation* where the expected value of Z is equal to its last observation Z_t: $Z^a_{t,t+1} = Z_t$.

The *rational expectation* of Z is:

$$Z^a_{t,t+1} = E(Z_{t+1}|I_t) \tag{B2.4.3}$$

where $E(Z_{t+1}|I_t)$ is the expected value of variable Z_{t+1} conditional on I_t, which represents the available information at time t, i.e., all relevant variables known by economic agents at the time of their decisions.

The difference from the previous formulation is that agents are supposed to make use, not only of the observed current and past values of Z, but also of the variables that determine Z. For example, in a floating-exchange-rate context they are supposed to know that a reduction of the exchange rate triggers a rise in domestic inflation, and therefore to regard the exchange rate as a leading indicator of future inflation.

With rational expectations, forecast errors (i.e., the difference between expected and actual values) are random. They cannot be forecast given the information available, because rational expectations are the best expectations that can be formulated on the basis of this information. This is especially relevant for foreign exchange and, more generally, for financial markets: One cannot expect to make a profit through making accurate forecasts of exchange rates, because the expectations rationally formed by market participants already incorporate available information on the determinants of the exchange rate.

In the special case where Z follows a *random walk**, i.e., it is the sum of random variables uncorrelated across time,

$$Z_{t,t+1} = Z_t + \varepsilon_{t+1} = Z_{t-k} + \varepsilon_{t-k+1} + \varepsilon_{t-k+2} + \ldots + \varepsilon_{t+1} \tag{2.4.4}$$

where ε_{t+1} is a random variable independent of all variables known at time t $\left(E\left(\varepsilon_{t+1}|I_t\right) = 0\right)$ then $E\left(Z_{t+1}|I_t\right) = Z_t$: The rational expectation of Z_t is equal to its static expectation. In plain English, this means that the best forecast of Z, given available information, is its last observed value. Such a model is empirically relevant for asset prices such as the exchange rate (see chapter 5). In general, however, the "true model" of the economy is much more complex. Agents use static expectations only when they lack a better method to forecast the future.

Muth's paper initiated what was named, somewhat bombastically, the *rational expectation revolution*, which had far-reaching consequences in all fields of economic policy. Think of workers preparing for a wage negotiation round. If they expect consumer prices to increase in the future, they will ask for higher wages to compensate future purchasing power losses. Of course, they cannot anticipate oil shocks or currency depreciations. However, if they know that a government policy is likely to

have inflationary effects, they will use this information and build it into their expectations, thereby bringing forward the inflation expected for the future.

The consequence is that in order to assess the impact of their decisions, governments have to take account of the expected reaction of economic agents. Economic policy is no longer the work of an engineer: It is the art of strategists interacting with other strategists. This is far more demanding.

b) Are expectations rational?

The rational expectation hypothesis was initially greeted with skepticism. Indeed, the assumption that the average economic agent has full knowledge of the functioning of the economy and is able to correctly anticipate all variables is an extreme one. It overlooks the simple fact that gathering and processing these data requires human capital and involves costs. The notion that households have enough economic culture, information, and computing skills to anticipate the effects of any economic policy on unemployment, inflation, or the public deficit defies intuition.

However, the alternative assumption that individuals do not at all use information available to them is not attractive either. And the rational expectation hypothesis does not require them to *know* all the laws of the economy, but only to *act* in accordance with them. Economic agents are akin to a character described in Robert Musil's novel *The Man Without Qualities*: Industrialist Arnheim does not know the laws of motion of the billiard ball but nevertheless knows how to play it:

> If I wished to state them theoretically, I should have to make use not only of the laws of mathematics and of the mechanics of rigid bodies, but also of the law of elasticity. I should have to know the coefficients of the material and what influence the temperature had. I should need the most delicate methods of measuring the co-ordination and graduation of my motor impulses. . . . My faculty of combination would have to be more rapid and more dependable than a slide-rule. . . . I should need to have all the qualities and to do all the things that I cannot possibly have or do. You are, I am sure, enough of a mathematician yourself to appreciate that it would take one a lifetime even to work out the course of an ordinary cannon-stroke in such a way. This is where our brain simply leaves us in the lurch! And yet I can go up to the billiard-table with a cigarette between my lips and a tune in my head, so to speak with my hat on, and, hardly bothering to survey the situation at all, take my cue to the ball and have the problem solved in a twinkling!
>
> Musil (1930, 1979)

In the same way, the Zimbabwean employee of the 2000s or his Bulgarian counterpart of the 1990s who anxiously watched the exchange rate to forecast future inflation did not necessarily know *why* a depreciation would trigger

inflation, but they had learned from experience that this was likely to happen. Households usually do not spend much time studying economic policy, but they have to do so when expectation errors would be costly for them. This is the case in extreme cases, such as hyperinflation (box 2.5) and more generally when a major policy change is expected, such as a swift fiscal retrenchment or a major tax reform.

For some other agents, the rational expectations hypothesis is a natural one. Banks and asset managers who operate in financial markets invest significant resources in economic research, in particular to forecast prices, interest rates, and exchange rates. Fed-watchers and European-Central-Bank-watchers are paid to gauge the next central bank decisions, and forward interest rates (as observed on future markets) actually track monetary policy decision quite accurately. It would be unrealistic to suppose that their expectations are naively backward-looking.

From a methodological standpoint, rational expectations merely impose a consistency constraint on model builders: It cannot be assumed that individuals make assumptions that contradict the model. They can also be seen as the limit on which expectations converge when individuals with initially adaptative expectations (box 2.4) accumulate knowledge on the functioning of the economy.[8]

Box 2.5 Rational Expectations in Action: The Bulgarian Currency Board

Sofia, Spring 1997. With a monthly inflation rate of 40%, Bulgaria was on the brink of hyperinflation and its currency, the lev, was in free fall. On 19 April, the Christian Democratic opposition won the elections. The new government confirmed its will to anchor the currency to the deutschemark through a currency board (a fixed exchange-rate regime, see chapter 5). On 1 July, the currency board was successfully introduced and inflation began to fall.

In June, just before monetary reform, an opinion poll asked Bulgarians citizens to assess future inflation depending on whether or not a currency board would be established. On average, their answer was that inflation would be 25% a year if a currency board was established, 50% if not. The prospect of monetary reform therefore had a major impact on expectations.

In the Bulgarian case, two conditions were conducive to such an expectation shift: High inflation meant that any expectation error by an individual agent could induce a significant economic cost; and the

8. Bayesian calculus is used to model this learning process, see Evans and Honkapohja (2001).

introduction of the currency board had been preceded by a national political debate, so that everyone knew what was at stake.

Source: Carlson and Valev (2001).

Summing up, rational expectations should be considered as a reference case, from which one can then depart to enrich the description of reality. One possible departure consists in recognizing that the information available to economic agents and the resources they can invest in its acquisition and treatment are heterogeneous. Another one is to abandon the rationality hypothesis and study in more detail the way agents form their judgment. This research avenue was opened by Daniel Kahneman and Amos Tversky at the junction of economics and cognitive sciences (see the Nobel Prize Lecture by Kahneman, 2002). Experimental economics nowadays constitutes a very active scientific field.

c) The Lucas critique

Pushing the reasoning further, Robert Lucas showed in a seminal 1976 paper that it is incorrect to use a macroeconometric model (see box 1.6 of chapter 1) to assess the consequences of systematic economic policy changes. For example, it is incorrect to rely on a standard simulation using such a model to evaluate the effects of moving from a monetary-targeting rule to an inflation-targeting rule or from a fixed to a floating exchange rate.[9] This is because the model's parameters have been estimated over the past: Systematic policy changes will be incorporated into the agents' expectations and will affect their behavior, of which the model is a representation (box 2.6). Lucas made a dominant and rapidly developing methodology shake to its foundations. Economic policy could no longer rest on an overly naive representation of the behavior of economic agents.

Box 2.6 The Lucas Critique

The Lucas critique is addressed to the use of macroeconometric models generally made of a large number of behavioral equations (consumption, investment, etc.) to assess the consequences of policy decisions. Typically, the model is of the type:

$$Y_t = h(X_t, Y_{t-1}, \varepsilon_t) \qquad \text{(B2.6.1)}$$

9. See chapters 4 and 5 for developments regarding monetary and exchange-rate rules.

where Y_t is a vector of variables representing the economy at time t, X_t is a vector of policy variables and ε_t is a vector of random shocks. The econometric estimation of h consists in summarizing the relationships between the variables Y and their determinants into a linear function H and a vector θ of parameters, so that:

$$Y_t = H(X_t, Y_{t-1}, \theta) + u_t \qquad \text{(B2.6.2)}$$

The value of θ is then derived from the observed historical data $(X_1 \ldots X_T)$ and $(Y_1 \ldots Y_T)$ so as to minimize a function of the vector of *estimation residual** $[Y - H(X, Y, \theta)]$. For example, the model parameters θ are chosen so as to minimize the sum of squared residuals.

When using the model for policy-evaluation purposes, the observed sequence of policies $(X_1 \ldots X_T)$ is replaced by an alternative one. For example, a higher sequence of public expenditures can be used to simulate the impact of a fiscal expansion.

In his paper, Lucas pointed out that this method makes sense only if the function H and the parameter vector θ are stable over time and do not depend in a systematic way on the policy sequence X_t. However, the agents' behavior depends on their expectations of the future values of the variables that affect their environment. H is stable only if the policy change does not affect these expectations. This may not be true and is generally not true for changes in the policy rules or the policy regime. As we shall see in chapter 5, expected inflation is not the same in a floating and in a fixed exchange-rate regime. Similarly, an investment equation estimated in a stable tax environment can be used to study the impact of improving order books, but not the impact of a permanent tax reform.

Any change in the economic policy framework thus modifies the very structure of the model, which cannot be considered invariant with respect to the phenomena it intends to study. The Lucas critique is not addressed to the choice of the model, i.e., the difference between the chosen model and the "true," unknown functioning of the economy. It is deeper in that it is based on the interaction between economic policy and the behavior of economic agents. It has sometimes been compared, in this respect, to Heisenberg's uncertainty principle.[10]

Research has addressed Lucas's objections. First, econometricians have done their best to rely on *micro-founded models**, i.e., models where private behaviors (consumption, investment) rely on explicit optimization with rational expectations (see box 1.6 of chapter 1). In those models, the "deep" parameters that determine the agent's long-term response to policy changes, such as the preference for the present or elasticities of substitution, are calibrated or, if estimated, are likely to be independent

10. This is a principle of quantum mechanics which states that the position and the speed of a particle cannot be observed simultaneously.

of the policy regime. Recursive resolution also allows expectations to be explicitly dependent on the models' results, which ensures consistency (this amounts to supposing that agents have the same knowledge of the economy as the econometrician who built the model). Such models make it possible to compare various economic policy regimes. The SIGMA model of the Federal Reserve (Erceg et al., 2005) and the GEM model of the International Monetary Fund (Botman et al., 2007) are examples of such models.

Second, the traditional macroeconometric approach was renewed by Christopher Sims' promotion of a constraint-free approach to the relationships between variables (Sims, 1980). This has led to the development of vector autoregressive (VAR) models (see box 1.4 in chapter 1) where the economy is modeled by a linear dynamic equation:

$$Y_t = \sum_{k=1}^{t} A_k Y_{t-k} + \varepsilon_t \qquad \text{(B2.6.3)}$$

where Y is a vector of n variables and A_k an (n, n) matrix of estimated coefficients. Unlike traditional models, VARs do not start from *a priori* restrictions on the value of the A_k coefficients. In particular, they do not determine *a priori* which variables are regarded as exogenous policy variables. This means that systematic policy reactions to shocks—for example, how the central bank responds to a rise in the inflation rate—are estimated in a similar way to the one used for private behavior. Though not immune from the Lucas critique, VARs thus have the advantage of embodying the interplay between private and policy players.

The two approaches are not incompatible: The VAR model can be estimated with constraints imposed on the coefficients (they are then called structural VARs). If these constraints are based purely on theoretical consistency, then they are not liable to the Lucas critique. For example, it can be assumed that monetary shocks do not have a long-term impact on output and prices (see the example in box 4.12 in chapter 4).

Not all empirical evaluations of economic policy are doomed by the Lucas critique. Macroeconometric models remain relevant to studying the effects of policy decisions that are nonpermanent or remain within the range of policy changes observed in the past. This, for example, applies to small-scale changes in public expenditures, tax rates or the interest rate. However, they cannot be used to evaluate the effects of a change in the *policy regime**, that is, of the principles and rules governing economic policy. For example, a model estimated in a floating-exchange-rate context could not be used to assess the effects of Slovakia joining the European Monetary Union on 1 January 2009, because this approach would have neglected the impact of the new monetary

regime on the agents' inflation expectations, and hence wage bargaining, and of financial markets integrated with the eurozone.

d) Implications for policy

The Lucas critique has contributed to making governments and central banks aware of the pitfalls of quantitative policy evaluations. By undermining confidence in those evaluations, it has contributed to weakening the technocratic approach to policy choices that prevailed in the 1970s. While evaluations with large-scale models are still carried out, they are used with greater caution, especially for substantial policy changes. They are mostly regarded as inputs into the policy process, alongside qualitative assessments or evaluations based on instruments that are robust against the Lucas critique, such as VARs or micro-founded models.

2.1.3 The limits of confidence

As explained in the previous section, rational expectations add complexity to the representation of the economy and of its interactions with economic policy. However, their impact goes beyond this mere technical difficulty. They may also directly hamper the effectiveness of public intervention.

a) Credibility

A compelling example with strong historical relevance (see chapter 4) deals with inflationary expectations. Assume a situation where wages are negotiated infrequently and where negotiated wages are rigid.[11] If wage-earners expect a price increase of 2% and have negotiated their wage increase accordingly, the government may aim at propping up inflation, say to 4%, so that the real wage (i.e., the nominal wage divided by the price of goods, see chapter 1) will be reduced *ex post*. Absent demand constraints, this should encourage job creation and lower unemployment, because selling prices rise more than unit labor costs. However, if individuals know in advance the government's plan, they will require a 4% wage increase in order to protect their purchasing power. The government will end up with a higher inflation (one speaks of an *inflation bias**) while real wages will remain unchanged. Strategic interaction between government and economic agents will result in inefficiency.

The issue at stake here is not government capture by special interests, or partisan politics ahead of the electoral cycle. It is the government's temptation to mislead private agents in the name of the general interest. By announcing that inflation will be 2%, then ensuring it is actually 4%, policymakers aim at reducing unemployment. But this seemingly virtuous lie is self-defeating.

11. The motives for price–wage rigidity are discussed in chapter 4.

The government may want to manipulate private agents, but it is in fact hostage to their expectations.[12]

Thanks to its simplicity, this argument formalized in 1983 by Robert Barro and David Gordon (the model is detailed in chapter 4, see box 4.8) exerted considerable influence on monetary policy thinking in the 1980s and 1990s. The same line of reasoning can apply to temptations regarding exchange-rate policy (make no announcement that you will devalue, and then take agents by surprise) or to the management of the public debt (issue long-term fixed-rate bonds, and then inflate away the public debt). It can also be extended to taxation. Imagine that a government announces that it will scrap taxes on fixed capital to encourage investment in its country, then reneges on its promise because it is socially optimal *ex post* to finance public goods by taxing capital. If companies anticipate this behavior, they will not invest at all.[13]

In all these examples, the problem arises from the lack of *credibility** of public intervention, i.e., governments do not succeed in convincing private agents that they will indeed behave in the way they have committed to. Reciprocally, a credible policy is all the more effective as it not only mechanically affects private behavior but also steers expectations. As we will see in chapter 4, this is particularly relevant for monetary policy, the effectiveness of which is based to a large extent on expectation management. An economy equipped with a credible central bank can better respond to inflationary shocks triggered by rises in the price of oil and raw materials because agents do not anticipate that these shocks will result in permanently higher inflation. Thanks to its credibility, the central bank can afford to let an oil-price shock trigger a one-off increase in the general price level without endangering its medium-term goals. This makes monetary policy more effective. In extreme cases, the expectation channel can actually be the only one to play a role. A case in point is Japanese monetary policy in the late 1990s and the early 2000s (box 2.7). This confirms Keynes's (1936) intuition that the "state of confidence" is the key variable in an economy prone to instability.

12. This reasoning should not be considered cynical and unrealistic on the grounds that governments do not and should not "mislead" the people. They do not have to actually do so, but only have the opportunity and temptation to do so. Thus, when governments do not even try to mislead the people, society may be penalized because they could try to do so.

13. The government can even nationalize private firms. The importance of property rights of foreign merchants was asserted in thirteenth-century England by King Edward I (Greif et al., 1994). Many developing countries have urged foreign companies to invest in their industry but have had a hard time convincing them that they would not nationalize the companies once the investment had been made. Property rights enforcement is a key condition for economic development (Djankov et al., 2003).

Box 2.7 Responsible and Irresponsible Credible Behavior

For a central bank, a key dimension of credibility is its ability to anchor inflation expectations. If its anti-inflationary stance is credible, short-term developments such as shocks to the price level or its own responses to them do not affect longer-term price expectations. This not only helps prevent inflationary spirals, but also gives the central bank more freedom of maneuver in setting interest rates. All modern central banks therefore attach great importance to remaining credible.

Whether a central bank is actually credible can be assessed from survey data. Many central banks run surveys of professional forecasters, such as the one presented for the European Central Bank (ECB) in figure B2.7.1. What is apparent in that graph is that five-year expectations are remarkably stable at a level that corresponds to the stated objective of the ECB. Shorter-term expectations are more volatile since they are affected by shocks and the responses to them, but volatility decreases as the expectation horizon lengthens. Judging from these data, the ECB has achieved a high degree of credibility.

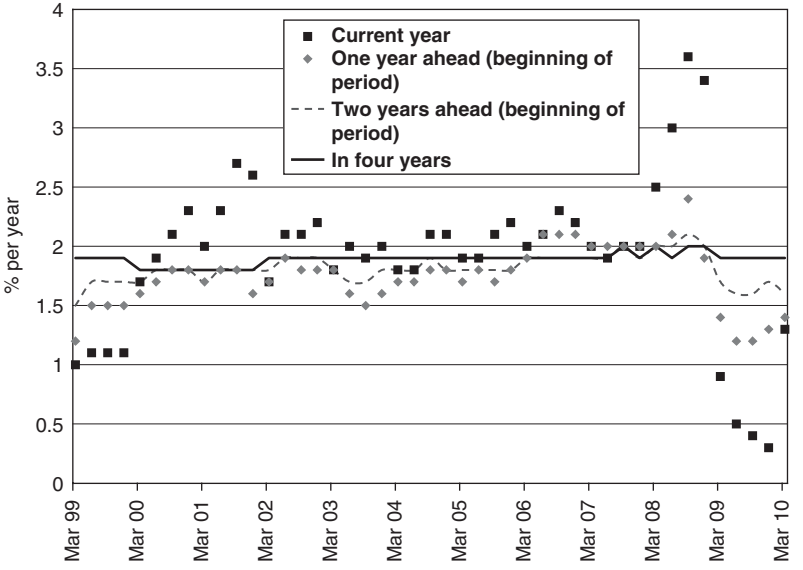


Figure B2.7.1 Inflation expectations in the euro area.
Source: ECB Survey of Professional Forecasters, April 2010.

It may sometimes (but more rarely) be important to be *credibly irresponsible.* In the late 1990s Japan was facing a deflationary crisis. The general level of prices was falling, generating positive real interest rates despite close-to-zero nominal interest rates. Traditional monetary policy was powerless. A deflationary spiral was looming.

In 1998, Paul Krugman suggested that if the central bank were able to generate positive inflation expectations, expected real interest rates (i.e., nominal interest rates minus expected inflation rates) would decrease, spurring investment and brushing away the risk of a deflationary spiral. He advised the Bank of Japan to "credibly promise to be irresponsible" and to enforce an inflationary policy. Krugman's unconventional proposal was initially regarded with a great deal of skepticism by the Japanese authorities, but they eventually agreed: The wish to promote inflation expectations contributed to the decision by the Bank of Japan in March 2001 to shift to a quantitative policy of monetary base expansion (see also chapter 4).

One year after, deflation fears resonated in the U.S.. In a landmark speech, a Federal Reserve Governor, Ben Bernanke (later to become the Fed President), discussed how a central bank could convince economic agents that in the event of a deflationary risk it would undertake with determination ... an inflationary policy (Bernanke, 2002). He had to heed his own advice in the wake of the financial crisis of 2007–09 in order to fight expectations of depression.

The concept of credibility has rapidly spread out of scholarly circles and has gained a wide audience in the public debate. For the sake of credibility, most countries have made their central banks independent and focused their mandate on fighting inflation, as we shall see in chapter 4. However, political leaders have not lived up to all the consequences of the credibility concept. German Chancellor Helmut Kohl famously promised in 1996 to halve unemployment by the year 2000 (it only went down from 8.5% to 7.2%), and the European governments that have pledged to bring their budgets into balance have a dismal record of broken commitments, a record that received a further and dramatic setback through the fiscal expansions in response to the 2007–09 financial crisis.

Unfulfilled promises undermine confidence in economic policy and hamper its effectiveness. That is why governments have increasingly put emphasis on acquiring and retaining credibility.

b) Moral hazard

We have seen that when expectations are rational, economic policy can become inefficient if the government seeks to mislead private agents. But the problem can be just as serious if it seeks to help them. *Moral hazard** is a well-known problem in insurance theory. By reducing the expected cost of future damages, insurance induces more risk-taking. Economic policy often provides insurance: Directly when the central bank assists banks that face a liquidity shortage or when the government rescues a distressed firm;

indirectly when stabilization policy prevents a recession. There is a tension between discouraging excessive risk-taking and helping involuntary victims of an accident.

Moral hazard is no theoretical curiosity. Compensating the victims of floods can encourage construction in areas likely to be flooded. *Ex post*, after the flood has occurred, there are very good reasons for the government to help displaced families. But *ex ante*, it should not provide free insurance. It should therefore either prohibit construction in areas susceptible to flooding or credibly claim that those who settle there do so at their own risk. Likewise, central banks generally aim at maintaining ambiguity and refrain from saying if and how they would provide liquidity assistance to distressed banks (see chapter 4). This was, for instance, the main argument put forward by the Bank of England in August 2007 to differentiate its policies from those of the ECB and refuse to inject large-scale liquidity. In the words of its Governor Mervyn King, "central banks cannot sensibly entertain such operations merely to restore the status quo ante. Rather, there must be strong grounds for believing that the absence of ex post insurance would lead to economic costs on a scale sufficient to ignore the moral hazard in the future".[14] Mervyn King had to change stance radically one month later when Northern Rock, a bank specializing in residential mortgages, fell short of collateral to borrow from the Bank of England[15] and created a run on the bank. Similarly, on 14 September 2008, the US Treasury refused to bail out the investment bank Lehman Brothers. Yet, faced with a major risk of a collapse of the financial system, the Treasury Secretary Henry Paulson proposed on 19 September a massive plan to buy distressed assets from banks and other financial institutions. These provide telling examples of policy-makers' time-inconsistent behaviors.

Another interesting example is the Russian financial crisis of 1998. Before it occurred, yields on Russian bonds implied a low-risk premium with respect to US Treasuries although the budget of the Russian federation was in dire straits. This was because the International Monetary Fund (IMF) was widely expected to help Russia repay its private creditors (it had already provided massive assistance to Mexico in 1994 and the Asian countries in 1997). This was a typical instance of moral hazard. But the decision by the IMF was to not add fresh money to the assistance already programmed. On 17 August 1998, the Russian government devalued the ruble and suspended debt repayment. Risk premiums increased sharply not only on Russian bonds but on all emerging market economies. As can be seen in figure 2.2, the annual yield spread between Brazilian and US Treasury bonds increased from 7.5% to almost 17% within a few weeks, while the economic situation of Brazil had not changed. The IMF decision on Russia thus acted as a signal.[16]

14. Mervyn King, *Turmoil in Financial Markets: What Can Central Banks Do? Note to the Treasury Select Committee of the House of Commons*, 12 September 2007.
15. Borrowing from central banks is explained in chapter 4.
16. Contagion of financial crises will be studied in chapter 5.

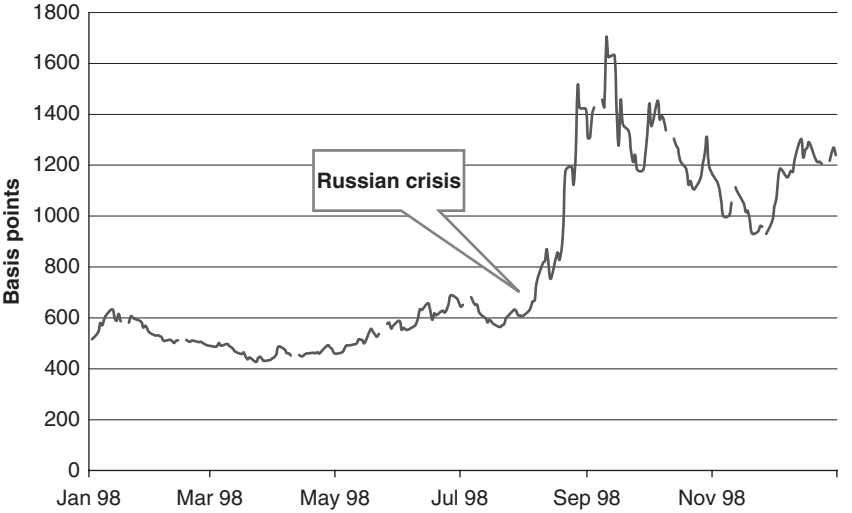


Figure 2.2 Yield spread between US and Brazilian Treasury bonds, 1998.
Source: JP Morgan, Emerging Market Bond Index, stripped spread.
Note: 1% is worth 100 basis points.

c) Time inconsistency

Lack of credibility and moral hazard are examples of what economists call *time inconsistency**: In both cases, the sequence of policy decisions that result from optimizing at each period does not constitute an optimal policy. In other words, *ex post* and *ex ante* optimality do not coincide. In the inflation bias example, it is optimal *ex ante* to announce low inflation, but it is optimal *ex post* to engineer an inflationary shock. In the flood example, it is optimal *ex ante* to announce that victims will not be compensated, but it is optimal *ex post* to minimize the consequences of the flood. In the 2008 Wall Street example, it was optimal *ex ante* to disavow any bank bail-out,[17] but it was indeed also optimal *ex post* to supply liquidities to the banking system to avoid a general banking panic that would have a devastating impact on the economy. The decision not to bail-out Lehman Brothers on 14 September 2008 had dire consequences on the evolution and contagion of the financial crisis.

The resulting inefficiency was established in a famous 1977 paper by Finn Kydland and Edward Prescott: They show that, except in specific cases, optimum policies are not consistent over time (box 2.8).

17. In this discussion, we ignore the collective benefits of insurance schemes. It may be desirable to introduce public insurance schemes in order to encourage risk-taking. For instance, there are guarantee funds for innovating firms and public systems of credit-insurance, notably for exporting firms.

Box 2.8 Time Inconsistency According to Kydland and Prescott (1977)

There are two periods. At each period $t = 1$ and $t = 2$, economic policy consists in choosing the value of the instrument X_t. The value taken by the target variable Y_t depends on the policies followed at the two periods (the fact that Y_1 depends not only on X_1 but also on X_2 reflects the influence of expectations). One has therefore:

$$Y_1 = G(X_1, X_2) \quad \text{and} \quad Y_2 = H(X_1, X_2) \tag{B2.8.1}$$

The decision-maker's objective is to maximize $U(Y_1, Y_2, X_1, X_2)$. In period 2, he or she can choose between:

- *Ex post* optimization: Select X_2 so as to maximize U, taking X_1 and Y_1 as given. This implies that:

$$\frac{\partial U}{\partial Y_2}\frac{\partial Y_2}{\partial X_2} + \frac{\partial U}{\partial X_2} = 0 \tag{B2.8.2}$$

- *Ex ante* optimization: Select X_2 so as to maximize U, taking X_1 as given but accounting for the fact that private agents' expectation of X_2 will influence Y_1. This implies:

$$\frac{\partial U}{\partial Y_2}\frac{\partial Y_2}{\partial X_2} + \frac{\partial U}{\partial X_2} + \frac{\partial Y_1}{\partial X_2}\left[\frac{\partial U}{\partial Y_1} + \frac{\partial U}{\partial Y_2}\frac{\partial Y_2}{\partial Y_1}\right] = 0 \tag{B2.8.3}$$

Ex post optimization coincides with *ex ante* optimization only if period 2 decisions do not affect the target variable in period 1 ($\partial Y_1/\partial X_2 = 0$). This happens if private expectations are backward-looking, or if changes in the value of Y_1 do not affect utility.Both assumptions are unlikely to be true under general circumstances. Hence, as a general rule, the government will be tempted to re-optimize in period 2, which will eventually lead it to renege on its initial announcement concerning X_2.

Kydland and Prescott's suggested response to time inconsistency is to banish *discretionary policies** that leave the policymaker free to decide which policy to follow at each point in time. In their view, economic policy should rather follow fixed *policy rules** that leave no or limited discretion to the policymaker, and economic policy evaluation should consist in comparing the performance over time of rules, not of individual decisions. This view of economic policy as a choice between alternative rules, rather than as a sequence of discretionary decisions, has been immensely influential.

d) Implications for policy

Criticisms based on credibility and moral hazard emphasize the intertemporal dimension of policy choices and the risks of adverse long-term effects of

seemingly optimal short-term decisions. They jointly lead to questioning of the traditional discretionary approach to policymaking and its call for leaving considerable latitude to the decision-maker.

Since the significance of the challenge began to be recognized in the 1970s, several strands of policy responses have been proposed and implemented. The first response has been rules-based policymaking, an approach introduced in 1979 in the US when the Federal Reserve endorsed a monetarist strategy focused on pre-announced quantitative targets. This mechanistic approach was abandoned in 1987 once inflation had been tamed and it had become clear that monetary aggregates provided poor guidance to monetary policy (see chapter 4), but it has become increasingly popular in the budgetary field (see chapter 3). Second, in the 1980s and the 1990s many governments in European and emerging countries "imported" credibility through committing to keeping the exchange rate stable vis-à-vis more credible and stable currencies (see chapter 5) or through scrapping altogether their currency. However, this has proved to be a risky strategy, as illustrated by a series of foreign exchange crises. Third, starting in the 1980s there has been a general move toward granting independence to central banks, as a way to ensure better credibility (see chapter 4). This approach has proved successful enough for the independent agency template to be proposed as a remedy to the pitfalls of discretionary fiscal policymaking. Finally, central banks themselves have introduced greater transparency in their objectives and decision-making procedures in order to convince the public that their deeds actually match their words. We shall return to all these techniques in the next chapters.

2.1.4 The limits of information

In the previous sections, we pointed out that governments could face limitations as regards the knowledge they have of the structure of the economy, but we have assumed that they had access to all available information. However, as already indicated in chapter 1, there can be other limitations that have to do with the strategic use of information by those who have access to it. The consequences of such informational asymmetries for private and public behavior have long remained underestimated, until economic theory started to explore them systematically in the late twentieth century. Joseph Stiglitz (2000, p. 1441) has argued that "the recognition that information is imperfect, that obtaining information can be costly, that there are important asymmetries of information, and that the extent of information asymmetries is affected by actions of firms and individuals is perhaps the most important innovation of 20th century economics." This also applies to economic policy.

a) Asymmetric information

In the traditional paradigm, government ministers are in command of a flawless administrative apparatus that provides them with accurate information

and seamlessly forwards their instructions from the top to the bottom of the bureaucracy—akin to the Soviet "Gosplan" (the central planning commission), which was supposed to determine the smallest details of the functioning of an economy of 300 million inhabitants. The lack of realism of the full-information assumption first emerged in the debate of the 1930s and 1940s between liberals and planners: The impossibility of gathering all information necessary to a centralized decision was turned into a powerful theoretical argument against planning, which Friedrich Hayek put at the heart of his criticism of central planning (Hayek, 1944).

Economists have brought into the picture imperfect information and the strategic behavior of government agencies and individual bureaucrats, and they have sobered up their conception of government. Indeed, when public or private agents have privileged information and use it strategically, the central decision-maker is in a situation of inferiority and his decisions are sub-optimal. When reporting to Moscow, Soviet companies systematically over-estimated their need for inputs (raw materials and machinery) and under-estimated their own productivity in order to meet their production targets more easily. The Gosplan did not have as much information as the company managers and could neither detect nor sanction this behavior. Similar problems arise in a host of situations. A telecommunications regulator may be tasked with controlling prices, but companies know technology and consumption patterns better than the regulator. When local authorities tender water supply contracts to private companies, they grant them exclusiveness of information on the technical state of the network and on water consumption. Health ministers would like to discourage over-consumption of health care, but doctors know patients' illnesses better.

These problems are not specific to the public sector. They are pervasive in market economies: For example, in the relationship between producers and consumers or between lenders and borrowers. An especially important case is that of the contractual relationship between what is generally called a *principal** (say, the shareholder of a firm or the manager within it) and one or several *agents** (say, entrepreneurs or employees). The principal, who delegates a task to the agent, does not have full information about the agent's capabilities and performance, and this generally leads to suboptimal situations (Laffont and Martimort, 2002)—a problem already discussed by Adam Smith in the case of the relationship between landowners and sharecroppers (or *metayers* in eighteenth century language).

b) Incentive contracts in the context of information asymmetry

The solution to this problem is to structure a principal–agent contract in a way that aligns the agent's interest with that of the principal and gives him incentives to reveal the information he has. This is what *contract theory**

is about.[18] Driven by expected profit, private agents—in particular companies—endeavor to transform their informational advantage into pecuniary revenue. In response, governments design contracts that give them incentives to reveal the information they hold. The telephone license auctions mentioned in chapter 1 are an example of a bidding mechanism aimed at revealing private information. Another example is the design of public procurements. Public contracts should be written in a way that ensures that it is not in the operator's interest to minimize technical problems (which would lower service quality), nor to exaggerate them (which could call for pecuniary compensation). This is generally done by conceding to the operator part of the operating revenue. Box 2.9 presents an example of an optimal procurement contract. The company's compensation is a convex and decreasing function of its production costs. This function can be understood as a "menu of contracts" offered to companies: Cost-effective companies are ready to assume a larger fraction of their costs, which they know are low, while less-effective companies want their costs to be supported by the contractor. By choosing a given contract, the company reveals otherwise private information on its cost structure. This is an example of *self-selection**.

Box 2.9 Optimal Public Procurement under Asymmetric Information

This model is inspired by Laffont and Tirole (1986) and Laffont (2000b). We consider a public procurement to a private contractor with unobservable operating costs. The goal is to devise a contract which allows high-cost companies to bid, without giving up all profits when the cost is low.

Cost Structure

The government contracts with a single, risk-neutral operator, to undertake a project which generates a social surplus Σ. The operating cost of the operator is:

$$C = \beta - e \qquad \text{(B2.9.1)}$$

β is exogenous and captures the technical characteristics of the company. e measures the cost-reduction effort. On top of the operating cost, there is an upstream cost $\psi(e)$ with $\psi(0) = 0$, $\psi' > 0$, $\psi'' > 0$ and $\psi''' > 0$, which measures the cost of achieving the effort level e through reorganization, training, knowledge management, etc. The government observes the operating cost C *ex post* (for example by auditing

18. See the survey by Holmström and Tirole (1989), and the textbook by Salanié (1997). Laffont and Martimort (2002) is a broader reference on the theory of incentives.

the company) but it does not observe its components β and e and even less the upstream cost $\psi(e)$. The company is refunded the operating cost C plus a flat fee t aimed at encouraging it to achieve the effort e. Public transfers are financed by taxes and there is an additional opportunity cost for the taxpayer, due to the government's own administrative costs and to the distortions induced by taxation. λ is the "production cost" of one dollar of government subsidy. The question asked by the government is how to set the flat fee t as a function of the observed cost C.

Surplus Analysis

The surplus of the operator is $S^f = t - \psi(e)$ and that of the government is $S^E = -(1+\lambda)(C+t)$. Σ denotes the (exogenous) surplus of other agents. The total social surplus generated by the project is thus:

$$W = \Sigma + S^f + S^E = \Sigma - (1+\lambda)(C + \psi(e)) - \lambda S^f \qquad \text{(B2.9.2)}$$

The contract between the government and the company is entirely summarized by $t(C)$: Knowing $t(C)$, the company can choose its effort level e so as to maximize S^f.

Optimal Contract under Imperfect Information

Under perfect information, the government observes β and e *ex ante*. It can at the same time maximize the social surplus W and capture the rent extracted by the company. However, in the general case, the government observes the total cost C but not the effort level e, which itself depends on the characteristics β of the company. Suppose that β is drawn randomly between $\underline{\beta}$ and $\bar{\beta}$ with a probability density f and a distribution function F:

$$F(x) = P(\beta < x) = \int_{\underline{\beta}}^{x} f(u)du \qquad \text{(B2.9.3)}$$

$e(\beta)$ and $C(\beta)$ depend on the realized value of β. The contract $t(C)$ has to meet two constraints:

- The *incentive constraint* $S^f(\beta) = \max_C\{t(C) - \psi(\beta - C)\}$: Knowing $t(C)$, the company chooses its effort level (and thus C) so as to maximize its surplus. A result from maximization theory known as *the envelope theorem* implies that $S^{f'}(\beta) = -\psi'(e(\beta))$. Contracts that induce a higher effort level e make the slope of $S^f(\beta)$ steeper and increase the revenue conceded to low-cost companies. The government faces a trade-off between setting the right incentive and limiting the rent extracted by the operator.

- The *participation, or individual rationality constraint*: $\forall \beta \in [\underline{\beta}, \bar{\beta}]$, $S^f(\beta) \geq 0$, meaning that the company has to be profitable. Since the revenue S^f is a decreasing function of β and there is an opportunity cost to spending public money, the constraint has to be saturated for the highest value of β and can thus be written: $S^f(\bar{\beta}) = 0$.

From these two constraints follows the relationship between the company surplus and its effort level:

$$S^f(\beta) = \int_{\beta}^{\bar{\beta}} \psi'(e(x))dx \tag{B2.9.4}$$

As for the government, it maximizes the expected total surplus:

$$Max\ EW(\beta) = \int_{\underline{\beta}}^{\bar{\beta}} \Big[\Sigma - (1+\lambda)(C(\beta) + \psi(e(\beta))) - \lambda S^f(\beta)\Big] f(\beta)d\beta \tag{B2.9.5}$$

The first-order condition reads:

$$\psi'(e(\beta)) = 1 - \frac{\lambda}{1+\lambda}\frac{F(\beta)}{f(\beta)}\psi''(e(\beta)) \tag{B2.9.6}$$

which makes it possible to derive the effort function $e(\beta)$, which is decreasing in β, and $t(C)$. The closed-form solution depends on the shapes of the function ψ and distribution f. It can be shown that in all cases, the optimum contract $t(C)$ is convex and decreasing in C (figure B2.9.1).

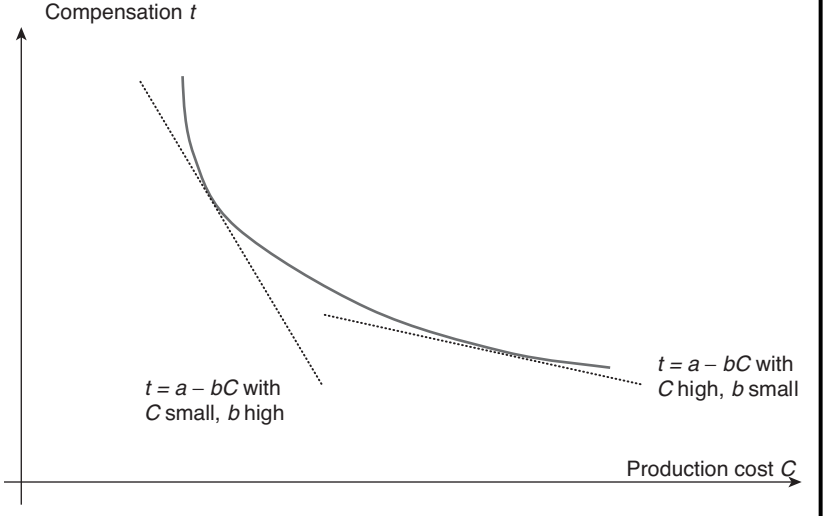


Figure B2.9.1 The optimal contract between the government and a contractor.

There is an intuitive interpretation of this result. In the vicinity of each cost level C, $t(C)$ is similar to a decreasing linear contract $t = a - bC$. The slope b represents the fraction of costs supported by the company. Since $t(C)$ is convex, b decreases the higher the cost level C. Inefficient companies wish a large proportion of their costs to be supported by the government. More-efficient ones are ready to support a larger fraction of costs, since they know they are low.

c) Implications for policy

This method has wide implications for public management, in areas such as public service delegation for infrastructure maintenance, waste disposal or water supply, public–private partnerships to build hospitals, schools, or prisons, or the regulation of natural monopolies such as rail infrastructures.

The same approach can be applied *within* the government. Rather than betting on the dedication of civil servants, incentive contracts can be devised so as to incite public employees to better achieve government objectives. This can be done through introducing performance-related compensation and promotion at the level of individuals, units, or departments. A famous, though seldom-implemented example is the so-called Walsh contract for central bankers, which makes the wage negatively dependent on the difference between the actual and the target inflation rate (see chapter 4).

2.1.5 The limits of benevolence

So far, we have not questioned the government's objective. It has been supposed to serve the general interest as defined in chapter 1 through a social welfare function. What the arguments outlined here underline are the limits to a government's capacity to act in an effective way when private agents behave strategically or in the presence of uncertainty.

The criticism of a government's capacity and willingness to serve the general interest is deeper and of a different nature. Building on earlier insights, modern research has called into question the far too naive vision of a well-informed and benevolent government that inspired normative economics and, in many countries, still constitutes the intellectual backbone of public service. The famously centralized French system is a case in point. As Jean-Jacques Laffont once commented: "The official and administrative system [...] rests on an idealized vision of political power and democratic life, on a general postulate of benevolence of politicians, the administration and all government officials and assimilated staff".[19] In Laffont's view, the notion that a politician's behavior can be described (as we did in chapter 1) as the maximization of a social

19. Laffont (2000a), p. 118 and p. 124, translated by the authors.

welfare function can be traced back to Jean-Jacques Rousseau, the French eighteenth-century philosopher, and to his vision of the government as a "frictionless device" and a mere "implementation instrument of the people's will," without a proper existence. From this perspective, today's representative government, which delegates policy implementation to bureaucrats, departs from this ideal and is a mere "technical artifice resulting from a purely material constraint."[20]

a) Why politicians may depart from the general interest

In addition to the informational dimension discussed in the previous section, five main, non-mutually-exclusive arguments have been mounted against Rousseau's paradigm.

First, politically accountable governments are *vulnerable to lack of credibility* and time inconsistency because exposure to opinion polls, short mandates, or the threat of losing a majority in parliament render investment in the build-up of a reputation difficult. They may therefore engage in policies that are not optimal from an intertemporal point of view.

Second, governments are exposed to *pressures from interest groups*. In most countries, the Agriculture Minister is as much the representative of farmers within the government as he or she is the voice of the government vis-à-vis the farmers. The advice he or she provides to the Prime Minister or the President is biased toward the interests of the sector. The Labor Minister is likewise sensitive to the arguments of trade unions, the Defense Minister to those of the military, and the Finance Minister to those of bankers.

The underlying problem is that citizens exhibit heterogeneous preferences. Each of them uses the political system so that the collective decision reflects, as much as possible, his or her own preferences. More than through the majority vote, this can be achieved through *lobbying** politicians and civil servants. Lobbying, which most countries nowadays recognize as a legitimate contribution to policy discussion, is usually intermediated by organizations known as *interest groups**, such as trade unions, consumer or environmental protection associations, industry representatives, community groups, etc.

Government money earmarked to particular interests (familiarly dubbed *pork barrel** in American English[21]) amounts to a tax levied on all taxpayers to the benefit of the few. Pork-barrel politics is an important dimension of the budgeting process in any country, as it is for the allocation of Europe's structural funds, which finance local infrastructure projects. It notably uses instruments such as campaign financing, media pressure, indoctrination, and corruption. However, it can also be more broadly understood as a political process that generates necessary and potentially welfare-improving trade-offs among various interest groups.

20. Rosanvallon (2000), p. 12 and p. 23.
21. The term originated in the pre-Civil-War period, when slaves were rewarded with salted pork.

The role of interest groups had long been known by sociologists and political scientists before it was acknowledged and modeled by economists. It was not until the early 1970s that George Stigler (1971) spoke of a *capture of the regulator** by the very interests he or she is responsible for supervising. Ever since, public economics has aimed at better identifying this risk and at defining how the regulator's mandate can be drafted in order to align his or her interests with the general interest.[22]

Third, governments are subject to *reelection* and are naturally motivated by it. The view that politicians are motivated only by the general interest from the first day of their mandate to the start of the next electoral campaign is overly naive. A government can act in an opportunistic way and seek re-election by lowering taxes just before a poll (at the risk of having to raise them later), by increasing its expenditures or by delaying difficult decisions. This type of behavior gives rise to a *political business cycle**.[23] In France, for example, the influence of the municipal electoral cycle (a six-year cycle) on local governments' investment is depicted in figure 2.3 for a sample of 58% of the communes. All things being equal, investment increases on average by 6% over the two years preceding a municipal election and falls by almost 5% over the two years following the election.

The simple political business cycle model rests on the assumption that citizens are not well-informed enough to decipher the politicians' tactic. However, a similar behavior can also emerge as regards public finance, retirements, or the environment, through making intertemporal choices that are systematically biased against the future generations that do not vote.

Fourth, governments can be *partisan* and, rather than serving the general interest, they may take measures that correspond to their prejudices or favor the majority that supports them. A reason for such behavior is that politicians are torn between what Max Weber (1919, 1978) called the "ethics of responsibility" and the "ethics of intention." They are not only accountable to the citizens at large, but also to their supporters and to those who share their beliefs.

Competition compounds the problem. Let us suppose, for example, that one of two competing political parties wishes to direct public investment toward defense, and the other one toward social housing. Knowing that if it loses power, priorities will change, the governing party, if doubtful of its re-election, will have a strong incentive to over-invest in its priority area and, at the same time, limit its successor's ability to spend through leaving it a high public debt. The more the country is divided and the more frequently power shifts between parties, the higher the public debt will be. The problem

22. Recognizing the existence of interest-group pressure is not an insult to civil servants' dedication to the general interest. It is only the recognition that it would be inefficient to put them in situations where their personal interests would not be aligned with their professional duty.

23. The expression *political business cycle* was introduced by William Nordhaus (1975. Empirical observations tend to confirm the existence of such a cycle. See for example Persson and Tabellini (2001).

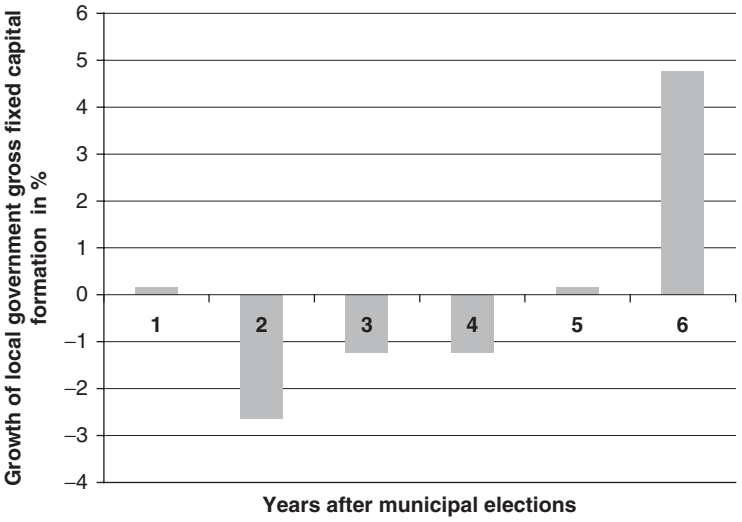


Figure 2.3 Electoral cycle and local investment in France.
Source: Besson (2002).
Note: Contribution of the municipal electoral cycle to gross fixed capital formation, averaged over 1965–2000.

here comes from each camp's conviction that its policies correspond to the general interest. Under those conditions, ideological division leads to excessive public spending and debt.[24] It can be confirmed empirically that public debt is positively correlated with the degree of political instability.

Fifth, *divisions between regions*, or *between ethnic or social groups*, may lead to inefficient spending. In such situations (which are often observed in newly created countries but can also arise in developed countries), each faction tries to extort from the government tangible benefits whose corresponding macroeconomic costs (higher public debt or inflation) will be distributed among the whole population. In this case, theory suggests that public spending will be too high, as well as public debt (in the event of debt financing) or inflation (in the event of monetization). There are many examples of such situations, in particular the impact of intercommunity tensions in the 1970s and 1980s on the Belgian public debt; the incapacity, in 2000–01, of the Argentine Federal State to get regions to contribute to sound public finance management; or the inflationary behavior of the former Soviet republics in the early 1990s, after the USSR had been dissolved but while the ruble remained. We will come back to these issues in chapter 3 by examining the consequences for public finance of "wars of attrition" on the distribution of the costs of a fiscal adjustment.

24. For formalization, see section 6 in Persson and Tabellini (1990).

b) Modeling politicians' behavior

Politicians' behavior has been modeled in several ways.[25] In the simplest theoretical models, politicians have no preferences of their own; their only objective is to be in power. Once elected, they seek to be re-elected.

It would seem that if politicians are only motivated by (re)election and voters are well-informed, decisions by politically motivated governments will coincide with the maximization of social welfare or with the decisions by the benevolent dictator of chapter 1. In fact, this is generally not the case.

The reason is the following: Majority vote gives a prominent role to the *median voter** (box 2.10). For instance, if left-wing and right-wing parties disagree on the level of government transfers, voters will choose the median level of transfers, i.e., half of the voters would like the level to be lower and half of them would like it to be higher. This is quite a logical outcome in a democracy. However, except under very specific assumptions, this does not coincide with either of the social choice objectives outlined in chapter 1. "Benthamian" choice would structure spending so as to maximize average welfare, while "Rawlsian" choice would concentrate transfers on the poorest.

Box 2.10 The Median Voter

The median voter model was introduced by Black (1948) and builds on the insights of Hotelling's (1929) model of competition. Suppose that voters' preferences can be represented along a single dimension, from "left" to "right" and that the government is elected by simple majority. Suppose furthermore that the competing parties' programs can be represented on the same axis. A voter will choose the party whose preferences are close to his or her own: Voters V_1 to V_4 will for example vote for candidate C_1 and voters V_5 to V_7 for candidate C_2 (figure B2.10.1).

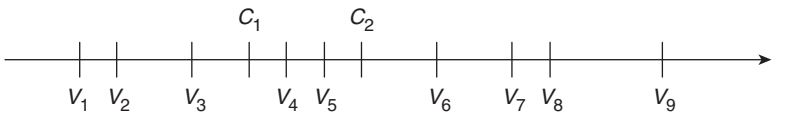


Figure B2.10.1 Preferences, votes, and the median voter.

If there are only two parties, one from the left and one from the right, each party will always capture the more extremist voters. V_1 for example has no choice but to vote for C_1, even though C_1 is much too centrist to its taste. Clearly, this implies that both parties have an incentive to tilt their programs toward the preferences of the center to capture as many voters

25. See Persson and Tabellini, 1999 and 2001, for a survey.

as necessary to reach power. This will lead them both to converge on the preference of the median voter V_5. This has two important consequences:

- The program of the winning party is aligned with the preferences of the median voter
- It is immaterial which party wins the election since both have converged on the same set of policies.

This result, known as the median voter theorem, mimics Hotelling's results on product differentiation by monopolistic producers.

There are clearly many simplifying assumptions in this model. Nevertheless, it captures some important features of decision by majority and remains widely used.

The model can also be used to analyze decision-making processes in collegiate bodies where each decision is subject to a vote, which is not the case within a government, where the President or Prime Minister has the last word. This is the case in the IMF, the European Union (EU), and the ECB. IMF directors represent groups of countries, whose vote is weighted according to their quota in the Fund capital. Voting rights at the EU Council of Ministers are also weighted. The ECB decides on the Eurozone single interest rate but inflation rates differ widely across member countries. The ECB Governing Council comprises six Executive Board members and as many national central bank governors as the number of member countries.[26] Suppose that each central bank governor favors an interest rate based on his own country's rate of inflation.[27] Majority vote would select the median interest rate, "Benthamian" choice would select the average rate (this is the behavior expected from the Executive Board members), and "Rawlsian" choice would aim at curbing inflation in the country where it is the highest. All three values would be different. A similar line of reasoning applies to the US Federal Reserve board, albeit with a smaller number of regional representatives.

The situation is even worse when there are multiple selection criteria. The *theory of social choice** has shown that in such situations, the aggregation of individual preferences may result in an impossibility. Even if each voter has a clear ranking among candidates, it is possible that a majority of them prefers A over B, another majority B over C, and a third majority C over A: This is the *Condorcet paradox**. This observation was formalized by Kenneth Arrow (1951): The *Arrow theorem** establishes that in the presence of at least three decision criteria, there is only one voting mechanism that does not make the

26. At a future point, the governors are committed to moving to a complex rotation system which will limit to 21 the number of central banks governors taking part in a vote at any point in time.
27. This is banned by the EU Treaty (chapter 4), but it is difficult to imagine that the governors do not take account of the preferences of their own home citizens.

relative ranking of two criteria dependent on the ranking of the other criteria. This mechanism is dictatorship.

More generally, a positive approach that explicitly takes into account the political process leads to envisaging how policy outcomes may depart from the optimum. A good example is a parliamentary decision on public spending and taxation. This type of decision is inherently political and results from the aggregation of a variety of preferences that are specific to regions, social groupings, or sectors. Careful analysis of this type of process leads to the conclusion that the budgetary decision may significantly depart from the social optimum (box 2.11). A significant body of recent research is devoted to understanding how political institutions—for example, a proportional versus a first-past-the-post voting system, or the allocation of agenda-setting powers between government and parliament—affect budgetary outcomes.

Box 2.11 The Political Economy of Public Spending

The level of public spending results from a political process, namely the vote on the budget in parliament. Depending on their situation and their partisan preferences, citizens hold different views on the budget's level and composition. Therefore, only a model that takes into account the political dimension of the decision process can explain what determines public spending.

To this end, Persson (1998) introduces a model where the utility u^{ji} of an individual j depends on his/her private consumption c^{ji} and on the consumption of a public good that is specific to a group i within society, g^i (we can think of regions, age groups, or sectors; there are at least three groups), to which he or she belongs:

$$u^{ji} = c^{ji} + \alpha^j H(g^i) \qquad j = 1, \ldots n;\ i = 1, \ldots K \qquad \text{(B2.11.1)}$$

α^j is the weight that individual j assigns to the public good. All individuals receive equal income y, pay taxes to finance the public good and consume the remainder:

$$c^{ji} = c^i = y - t^i \qquad \text{(B2.11.2)}$$

First, assume that both taxes and public-good provision are group-specific and decided within groups (think about local taxes that finance local infrastructure). Assuming $E(\alpha^j) = 1$ in all groups (same "average" preference), then the level of public-good provision is identical across groups:

$$t^i = g^i = H_g^{-1}(1) \qquad i = 1, \ldots K \qquad \text{(B2.11.3)}$$

Alternatively, assume that public-good provision is still group-specific, but that its financing falls equally on all citizens, and that the decision

is centralized. In this case, there is a negative externality stemming from the needs to finance other groups' public goods without benefiting from them:

$$c^{ji} = c^i = y - t = y - \frac{\Sigma g^i}{K} \quad j = 1, \ldots n;\ i = 1, \ldots K \qquad \text{(B2.11.4)}$$

If a subset of the groups is somehow able to capture the decision, it can twist it in its favor, which results in excess spending on some public goods. This corresponds to the familiar pork-barrel politics. Grossman and Helpman (1994), for example, have studied the effects of lobbying in a trade-policy setting.

The interesting issue is how majority decision on spending affects the outcome. Suppose now that this decision rests with a parliament where each group (in this case locality) has one representative. Suppose also that preferences are exogenous. Each representative tries to maximize the utility for her constituency l, which is:

$$U^l = y - \frac{\Sigma g^i}{K} + \alpha^l H(g^l) \qquad \text{(B2.11.5)}$$

Finally, suppose that one of the representatives has agenda-setting power, i.e., that she is able to present a budget to the vote of her fellow members of parliament, and that in the absence of a positive vote, taxes and public spending are set at zero. The other representatives will approve the budget if it improves the situation in comparison to this default solution. So the agenda-setter knows that she needs to assemble a majority of at least half the members of parliament and she sets itself the goal of maximizing (B2.11.5) subject to this constraint. The outcome is a budget where spending is high in the locality of the agenda-setter, intermediate in the localities whose representatives participate in the coalition, and nil elsewhere. Assuming the agenda-setter builds a coalition at minimum cost for her, this coalition will consist in those representatives whose localities are characterized by the highest α^l, since they are the cheapest to buy off.

This outcome is likely to be socially inefficient because spending is too high in the agenda setter's locality and too low in localities outside the majority. The model is clearly oversimplified since it does not take into account partisan preferences and starts from an excessively rough representation of the decision. Nevertheless, it captures an important insight.

c) Implications for policy

Taking on board the political dimension should not result in sheer skepticism toward economic policy. It merely acknowledges that political institutions shape economic outcomes, and that they should be framed so that the outcome

of political processes corresponds to the general interest. In this respect, the political economy approach can help in designing and adopting policy institutions that are conducive to socially desirable outcomes. The process of institutional selection that took place in the 1980s and the 1990s for central banks and led nearly all of them to become independent can therefore to some extent be replicated for budgetary or regulatory institutions.

Furthermore, public decision-makers cannot ignore the risk that government employees themselves have a biased vision of the general interest. Unlike politicians, civil servants are not motivated by the desire to be re-elected, but by career concern. Civil servants' incentives are thus not aligned with voters' preferences unless their mandate is clear and verifiable. They can also be motivated by the prospect of future employment in the private sector and in some cases by corruption. All advanced countries have experience with corruption in public procurement, and they have put in place codes of ethics to regulate civil servants' relationships with the private sector. The risk of corruption is even higher in low-income countries, where civil servants are badly paid. The structure of governmental institutions and of the political process are important determinants of corruption (Shleifer and Vishny, 1993).

2.1.6 The Policy Responses

Now that we are aware of the various limits of economic policymaking and the necessity of creating adequate institutions to address these limits, it is time to examine how economic policy decisions are made in practice.

The last quarter of the twentieth century witnessed the emergence of two major governance technologies: First, the creation and development of a number of specialized agencies or institutions with independent policymaking or monitoring power; second, a significantly greater reliance on rules that constrain the behavior of policy authorities.

a) Delegation to independent agencies

The recourse to independent authorities that act independently on behalf of the parliament or the government has old roots. The *Bank of England* was created in 1694 (though it was only made independent three centuries later), the US *Interstate Commerce Commission* was born in 1887, the *Securities and Exchange Commission* (SEC) was created under Roosevelt in 1934, and in Germany the *Bundesbank* was introduced in 1947 and the federal office for anti-trust (*Bundeskartellamt*) in 1958. The granting of policy powers to independent agencies has, however, markedly accelerated in the last quarter of the twentieth century, especially in Europe and more recently in emerging countries.[28] Delegation to an independent body has even become

28. In the US, such agencies have existed for a long time, but their field of competence has hardly expanded.

the dominant model for central banking, competition, and sectoral regulation. According to Gilardi (2005), the proportion of European countries equipped with independent agencies for competition, financial regulation, and the regulation of telecoms was below 10% in 1960 and below 20% in 1980, but it had reached 90% in 2000.

Delegation does not take place only within countries. The EU also provides examples of various such authorities (like the *European Central Bank* or the *European Commission* in its function as a guardian of competition) or international committees formed by the regulators themselves (like the *Committee of European Securities Regulators*—CESR). Private regulatory bodies have also emerged internationally, such as the *International Accounting Standards Board*, which sets international accounting standards.

The proliferation of independent authorities has been criticized by numerous politicians from the left as well as from the right. It actually raises two main questions.

The first question is why and when it is preferable to remove certain fields of public decision from direct political influence. In a democracy, these institutions perform under a mandate given by the legislator, who keeps both the responsibility for defining and monitoring the mandate and the option to withdraw its delegation. Yet decisions on a case-by-case basis do escape control by the executive and, in some cases, it has been deemed preferable to deprive even the legislator of the right to amend the mandate, by anchoring it in a legal order higher than the law, as has been the case for the European Central Bank, whose independence is embedded in the Maastricht Treaty, and which can be changed only by unanimity of the 27 EU members.

The second question is how to conduct economic policy in a system where policy instruments are in the hands of independent bodies that may or may not coordinate with each other. For example, in 2007 in the UK, responsibility for financial stability was shared between the Treasury, the Bank of England, and the Financial Services Authority (see chapter 4). The three institutions (a government department and two independent agencies) were supposed to coordinate according to a memorandum of understanding. However, the run on a bank, Northern Rock (see above) exposed flaws in the system and triggered a debate about the wisdom of dividing up responsibilities between three different institutions. This is an example of the coordination difficulties raised by the delegation model.

When should, in a democracy, a decision be delegated to a technocratic body rather than to a government accountable before parliament? Why, for example, delegate the management of the currency to an independent central bank, and not that of the national budget and of taxes? Why establish competition authorities? Political science has for a long time been interested in these questions, but it is only since the 1990s that they have received systematic treatment in economic theory.

Contrary to perceptions, economic theory does not recommend an across-the-board delegation of responsibilities to nonelected authorities, nor does

it recommend in principle restricting the scope of democratic choice.[29] One can regret, with former Fed Vice-Chairman and Princeton Professor Alan Blinder, that the government is "too political,"[30] and yet agree with his former colleague Joseph Stiglitz, Chairman of President Clinton's Council of Economic Advisers, that technocratic bodies are not political enough.[31] Nonelected bodies are subject to failures that are symmetrical to those of governments: Behavioral rigidities, insensitivity to the society's expectations, inability to trade off between objectives, lack of legitimacy to deal with decisions that involve a distributional dimension. . . . As noted by Alesina and Tabellini (2007), foreign policy is also vulnerable to credibility and time inconsistency problems, yet nobody suggests that it should be delegated to an agency, notably because objectives and actions need to be re-assessed constantly and cannot be framed within a consistent and stable mandate.

Political and technocratic decisions are thus two imperfect methods of governance. One needs criteria to guide decisions to assign specific responsibilities to technocratic bodies—of course under a mandate defined and monitored by the legislator. Modeling the technocrat's and the politician's behavior (Maskin and Tirole, 2004; Alesina and Tabellini, 2007) leads to several general insights (cf. box 2.12), which need to be supplemented by judgment. Technocratic decision appears preferable when:

1. The matter is very technical;
2. Social preferences are stable and performance criteria are welldefined;
3. The decisions in question and their effects are not easily observable by voters;
4. The decisions are highly vulnerable to time inconsistency;
5. The decisions have a limited impact on income distribution within generations;

29. On this debate, see Fitoussi (2002).

30. Alan Blinder was successively member of the Council of Economic Advisers under the Clinton presidency and then Vice-Chairman of the Federal Reserve. " . . . life at the White House is fastpaced, exhilarating, and, of necessity, highly political. Policy discussions may begin with the merits ('Which option is best for the American public?'), but the debate quickly turns to such cosmic questions as whether the chair of the relevant congressional subcommittee would support the policy, which interest groups would be for and against it, what the 'message' would be, and how that message would play in Peoria." He then evokes the Federal Reserve, where, he notes, the reverse occurs. Blinder (1997, p. 117).

31. "If, as we have argued, there are alternative economic policies, and if these alternatives affect different groups differently, then it matters a great deal who makes decisions, and how those decisions are made. If there is an unemployment/inflation trade-off, and if workers care more about unemployment, while financial markets care more about the erosion of the value of their nominal assets with inflation, then workers and financial markets will see the trade-off in different lights; entrusting the decision about monetary policy to an independent central bank controlled by financial interests, or mandating that the central bank focus only on inflation, makes it more likely that the outcomes will accord with financial interests, rather than the interests of workers." Stiglitz (2003, p. 27).

6. The decisions significantly affect the distribution of income between generations;
7. The decisions do not involve trade-off between incompatible objectives;
8. The decisions entail benefits (or costs) to groups that are likely to be involved in political lobbying.

Of course, no economic policy issue completely meets the eight criteria, but they provide a useful analytical grid. For instance, monetary policy meets all the criteria except the seventh (at least in the short term, raising interest rates will slow down inflation and simultaneously increase unemployment) and perhaps the fifth (a drop in interest rates redistributes interest income from *rentiers* to indebted households and firms). However, the weighting of the objectives can be specified once and for all in the statute of the central bank (this point will be discussed in chapter 4). As for fiscal policy, it does not satisfy criteria 2, 3, 5, and 7. These are compelling reasons to keep fiscal policy within the realm of political decision-making.

Box 2.12 Technocrats or Politicians: Who Should Decide?

Eric Maskin and Jean Tirole (2004) on the one hand, and Alberto Alesina and Guido Tabellini (2007) on the other, have studied the choice between two governance regimes—by an independent agency or by political government—in a context of information asymmetry. The choice between a "technocratic" contract and a "political" one depends on the relative performance of the technocrat and the politician, given their respective incentives.

For Maskin and Tirole, the problem lies with the information the electorate has on the stakes involved in economic decisions. They use a two-period model. In each period, two decisions are possible, one of which corresponds to the social optimum. Voters are initially uncertain about which is the better policy. At the end of the initial period, however, they can, with a probability q, discover it (but they remain in uncertainty with probability $1 - q$).

Voters delegate the decision to a policymaker—an appointed technocrat (a "judge") or an elected officer (a "politician")—who is informed about the likely outcome of alternative policies, but who also pursues his own preference, which can differ from that of the voters. For example, voters do not know whether priority should be given to stimulating growth or to fighting inflation. They can delegate this choice, but run the risk that their delegate is either too strict or too lax, in comparison to their own preferences.

- Once named, the *technocrat* chooses the decision that he or she considers good, without consideration for the voters' preferences.

- The *politician* seeks to be renewed at the end of period 1, which can encourage him or her to behave in a demagogic way: If voters are mistaken about the nature of the good decision, the politician may decide to take the wrong decision in order to please the electorate and ensure his or her re-election. But she or he can also speculate that the electorate will learn what the good decision was, and reward him or her for having had the courage to confront opposed public opinion.

Maskin and Tirole find that the technocratic contract is preferable to the political contract when the probability q that the electorate will discover what the good decision was is low. In this case, the politician is unlikely to be rewarded for having taken the good decision and will prefer following the voters in their potential error. On the other hand, if voters acquire information with time, delegation to the elected politician is preferable to delegation to a technocrat (who presents the risk that she or he follows her or his own preferences, which may differ from the social optimum).

This model suggests that one should delegate to technocrats in areas where the electorate is poorly informed and unlikely to acquire information (for example, when the matter is too technical or of insufficient direct importance for citizens to invest in the acquisition of information).

Alesina and Tabellini emphasize motivations. For them, delegation to a technocrat or to a politician are two forms of contract concluded by a sovereign people. They thus follow the line of Laffont-Tirole (see box 2.9). Talents, effort, and outcome are in the two cases connected by a relation:

$$Y = \theta + e + \varepsilon \qquad \text{(B2.12.1)}$$

where Y is the outcome, θ a random variable representing talent, e the effort, and ε a random error term (assumed to be white noise). The question is then to know which contract, technocratic or political, provides the greatest incentive to effort, since neither talent nor effort can be directly observed.

The technocrat chooses his or her effort level e in order to maximize his or her utility function, which is the difference between reward $R^T(e)$ and the cost of effort $\psi(e)$. His or her reward is the expectation of his perceived talent, given the expected outcome Y:

$$R^T(e) = E\left[E(\theta|Y)\right] = E\left[E(Y - e^a - \varepsilon|Y)\right] \qquad \text{(B2.12.2)}$$

where e^a is the effort perceived by the public (at equilibrium $e^a = e$). The politician chooses in the same way his or her effort level, but his or her

reward is re-election, which depends on the probability that the result Y exceeds a threshold W.

$$R^P(e) = \Pr[Y \geq W] = 1 - \Pr[\theta \leq W - e - \varepsilon] \qquad \text{(B2.12.3)}$$

Each one of these two contracts therefore leads to a level of effort that results from optimization behavior by the agent having received the delegation. The first-order conditions yield the optimum level of effort in each case. Denoting by σ_θ^2 the perceived variance of talent across policymakers and by σ_ε^2 the variance of white noise, we have:

Technocrat:

$$\frac{\partial \Psi(e)}{\partial e} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2} \qquad \text{(B2.12.4)}$$

Politician:

$$\frac{\partial \Psi(e)}{\partial e} = \frac{1}{\sqrt{\sigma_\theta^2 + \sigma_\varepsilon^2}\sqrt{\pi}} \qquad \text{(B2.12.5)}$$

where the marginal cost of effort is an increasing function of effort ($\partial \Psi(e)/\partial e > 0$, $\partial^2 \Psi(e)/\partial e^2 > 0$). The main results are:

- The presence of noise reduces the level of effort in both cases: The higher the variance of ε, the less clear the relation between effort and performance, and the weaker the incentive to effort. Neither contract outperforms the other from this point of view.
- The variance of talent raises the technocrat's level of effort but it reduces that of the politician: When talent is uncertain, the technocratic contract is preferable, because the incentive to demonstrate competence is stronger. This suggests that it is preferable to delegate to technocrats jobs for which the dispersion of unobservable competences is large.

The same model can be used to choose to whom to entrust a responsibility that involves a trade-off between two objectives (or two alternative tasks on which the effort has to be allocated), when there is uncertainty about the preferences of the electorate. In this case, the allocation of the technocrat's efforts is specified *ex ante* and it is held there because it is the best means of showing its competence. On the other hand, the politician shows flexibility and adapts to changes in the electorate's preferences. The political contract is therefore more adapted.

In the same vein, the technocratic contract is shown to be preferable in the event of time inconsistency, while the political contract is superior when it is necessary to compensate the losers.

Beyond conventional wisdom and some disenchantment with politics, the tendency to assign certain fields of decision to independent agencies can be interpreted as reflecting:

- The increased technical complexity of a number of decisions, for example as regards sectoral or financial regulation, and in areas (e.g., risk prevention) where public decision relies heavily on scientific expertise.
- The judicial nature of some decisions, as regards, for example, merger control, the regulation of competition, or the enforcement of sanitary standards.
- The wish to constrain the policymaker's objective function and eliminate trade-offs with other objectives. This, for example, is the case with decisions that concern public health and safety, where public opinion does not regard any trade-off with economic or financial objectives as legitimate, even though such trade-offs do arise; or with monetary policy, where (some) countries have chosen to limit the trade-off between inflation and unemployment.
- The rising importance of intertemporal concerns. In a context of weaker productivity gains and of demographic decline, expectations of future income depend less on growth prospects and more on inflation. The independence of central banks or the success of sovereign wealth funds can thus be interpreted as guarantees given to savers that the wealth that they accumulate will be protected.
- An integrated global economy without a global government. In the absence of *ex ante* political legitimacy, international governance tends to rely on technocratic institutions in order to create, through the demonstration of its effectiveness, the conditions of *ex post* legitimacy.

That being said, the choice between political and technocratic governance is less clear-cut than it appears. Intermediate formulas do exist, like those in which elected officials choose the objectives and assign the responsibility for implementation to technocratic bodies that are granted operational independence. The relation between the UK Chancellor of the Exchequer and the governor of the Bank of England with respect to monetary policy provides an example (cf. chapter 4).

b) Policy rules

There has been a long-standing debate over public policy as to whether government decisions should abide by rules or be able to react on a case-by-case, results-based, optimizing basis. Rules are prescriptions for policymakers and other economic agents that are stable across time and therefore commit policymaking and private behavior for the future, even though they may be explicitly contingent on states of nature. Their role has notably received much attention in the field of regulation, with an ongoing trade-off between rules-based and principles-based regulation that emerged after the 2001 Enron

scandal and was revived in the wake of the 2007–09 crisis. Regulatory rules are often complex; as a result, monitoring their implementation is difficult. They also always present loopholes that can be exploited. Conversely, principles-based (risk-focused) regulations allow more discretion and may be less transparent, but under a strong, independent regulator can deliver results that conform better to a set of social objectives embodied in such principles.[32]

The debate about rules versus discretion, in the area of macroeconomic and especially monetary policy has been of a different nature. The argument for rules has evolved over time, from a focus on the lack of knowledge of policymakers to a focus on credibility and the time inconsistency of optimal policies. Governance by rules originates in the lessons drawn from the literature on economic policy evaluation (see section 2.1.2) and on time inconsistency (see section 2.1.3). Robert Lucas's critique of traditional policy evaluation led him to advocate comparing policy *rules* rather than policy *acts*: His main point was that only the results of rules can be rigorously compared (Lucas, 1976). Finn Kydland's and Edward Prescott's preference for rules over discretion rested on a different argument, namely that "selecting the decision which is best, given the current situation, [. . .] either results in consistent but suboptimal planning or in economic instability" (Kydland and Prescott, 1977, p. 487).

Rules were first tried—with limited success—with monetary policy. In the late 1970s and the early 1980s, the Federal Reserve briefly endorsed them when it adopted a strategy based on quantitative targets for monetary aggregates. The UK also implemented a similar strategy. Both experiments were discontinued after a few years. However, the rules-based approach to policy was revived in the 1990s when a growing number of central banks adopted explicit *inflation-targeting** strategies (see chapter 4). This approach consists in setting a target for inflation and in committing the central bank to following a course that ensures that future inflation (conditional on available information) is consistent with the prescribed objective. The complication here is that the central bank cannot commit to reaching a result because inflation depends on the occurrence of shocks (for example, shocks to the prices of oil and raw materials) that are beyond its control. However, it can commit to ensuring that forecast inflation remains under control and that forecasts are based on transparent and unbiased methods. Note that in this case the rules-based approach is used as a complement to, not substitute for, delegation to an independent authority.

In the budgetary field, rules were introduced later, but nowadays many countries, especially in Europe, have defined policy rules such as the European "Stability and Growth Pact" and the UK "Code for Fiscal Stability" (chapter 3). Their aim is to enforce responsible fiscal behavior over the medium term, while

32. Britain's Financial Services Authority (FSA) provides an example of a principles-based and risk-focused regulator. For the recent debate concerning rules-based and principles-based regulation in the US, see for example Bernanke (2007) and Wallison (2007).

leaving room for short-term stabilization. Those rules, however, have had limited success. This is most notable in the case of the European Stability and Growth Pact, which has been successful for some countries (Germany, Spain, Finland) but much less so for others (France, Italy, Portugal). Ownership of European rules by national governments and parliaments remains an open issue. By March 2010, due to the 2007–09 financial and economic crisis, 20 EU countries were under ongoing excessive deficit procedure and had received under that procedure recommendations from the Council to adopt corrective measures within set deadlines.

There also exist exchange-rate-policy rules such as currency boards and crawling pegs (chapter 5). They were widely used in the 1980s to anchor price expectations and demonstrate a government's commitment to price stability by attaching a highly visible political price to the option of inflating problems away. However, for countries outside a monetary union and without the prospect of joining one, exchange-rate regimes have evolved in the direction of increased flexibility.

Rules nowadays are less rigid than envisaged in the early monetarist writings, and they aim at combining medium-term discipline with a degree of discretion. This is being done by defining an explicit policy strategy that is followed unless unexpected developments lead to departure from it. In the latter case, policymakers need to explain why they have chosen to do so. This especially has advantages in the presence of Knightian uncertainty as defined in section 2.1.1, as policymakers in this case need to retain flexibility. In the words of Mervyn King (2004), the Governor of the Bank of England, "the ideal is a framework that will implement what we currently believe to be the optimal monetary-policy strategy and will deviate from that only if collectively we change our view about what that strategy should be." Such an approach is often called *constrained discretion** and it serves as a reference for several policy institutions, including the US Federal Reserve.[33]

2.2 Living with Interdependence

In the previous section, we have outlined the main limits to the traditional representation of economic policy decisions. So far we have not questioned the policymaker's ability to exercise responsibility for decisions with a bearing on her or his country. Reality, however, is increasingly distant from this single-country representation. Policymakers need to take into account the cross-border implications of their decisions and their interactions with other governments. A growing number of rules that constrain national policy choices are set at the global level. International institutions are entrusted

33. Note that Mervyn King speaks of strategies, not individual decisions—otherwise he would be giving an almost exact definition of time inconsistency.

with the responsibility for ensuring the consistency of national policies with international agreements, and also for conducting certain policies on behalf of the international community. Important policy competencies have also been transferred to regional institutions—the most notable being the European Union—or devolved to sub-national entities. This severely challenges the simple assumption that economic policy is exclusively or primarily conducted at the level of the nation-state.

A few examples will illustrate the degree to which interdependence matters:

- On 22 October 2007, Seattle-based Microsoft Corporation announced that it would cease challenging a decision by the European Commission requesting it to offer to competing networking software companies the information necessary to interact fully with Microsoft-operated desktops and servers. This was the conclusion of a procedure initiated in the late 1990s which had led to the fining of Microsoft, and against which the US software giant had in vain introduced an appeal. A few days later, on 26 October, the *Wall Street Journal* posted an editorial accusing the EU of "regulatory imperialism."
- On 12 December 2007 the European Central Bank announced that, in agreement with the US Federal Reserve, it would start offering short-term US dollar funding to banks in the euro area. This unprecedented agreement was a response to a growing shortage of liquidity in the money markets and the inability of some European banks to get access to US dollars through normal interbank lending.
- The set of prudential rules that are imposed on international banks was first defined in a 1988 decision of the *Basel Committee**[34], which gathers the central bank governors of the main developed countries. The agreement, notorious for its capital adequacy ratio (the "Cooke ratio"), has been written into domestic law and implemented on a country-by-country basis by the national supervisors. It was revised in 2006 and in 2010.
- The World Trade Organization (WTO) created in 1995 has responsibility for settling disputes between member countries on the basis of multilateral trade agreements. In a first step, the WTO's Dispute Settlement Body creates a panel to examine the dispute. In a second step, an Appellate Body may decide on cases that remain unsolved after a panel has reported. At the end of 2007 the panels had examined 132 bilateral disputes and the Appellate Body had decided on 84 cases brought to appeal by the parties involved.

34. The Basel Committee is a forum for cooperation on banking supervision whose secretariat is hosted at the Bank for International Settlements in Basel (www.bis.org). It was established in 1974 by the central bank governors of the G10. It has introduced in 1988 a capital measurement system (the Basel Capital Accord, introducing the so-called Cooke ratio), issued a revised capital adequacy framework in 2004 (also known as Basel II), and developed "core principles for effective banking supervision".

- According to the *Treaty of Lisbon** that was signed in 2007 by European governments and entered into force on 1 December 2009, the EU has exclusive competence (meaning that it has taken over competence from the member states) in the fields of customs union, the common commercial policy, competition rules necessary for the functioning of the internal market, and monetary policy (for the members of the euro area). Competences are shared between the EU and the member states in many other areas. When legislations conflict, the Lisbon Treaty states that EU law has precedence over national law.

2.2.1 The rise of interdependence

Interdependence is not easy to measure. One of the strongest forms of interdependence nowadays arises from the effect of each individual country's decisions on the global climate, yet this takes place without any cross-country trading or investment. Nevertheless, international flows in products, capital, labor, and technology, as well as cross-border holdings of productive and financial assets, provide a rough measure of international economic integration. Figure 2.4 illustrates the rise of interdependences: From the mid-1960s to the mid-2000s, the share of exports (or imports) in G7 countries' GDPs rose from 13 to 25%. Trade openness now significantly exceeds levels reached in 1913 at the end of the first phase of globalization. The rise of gross

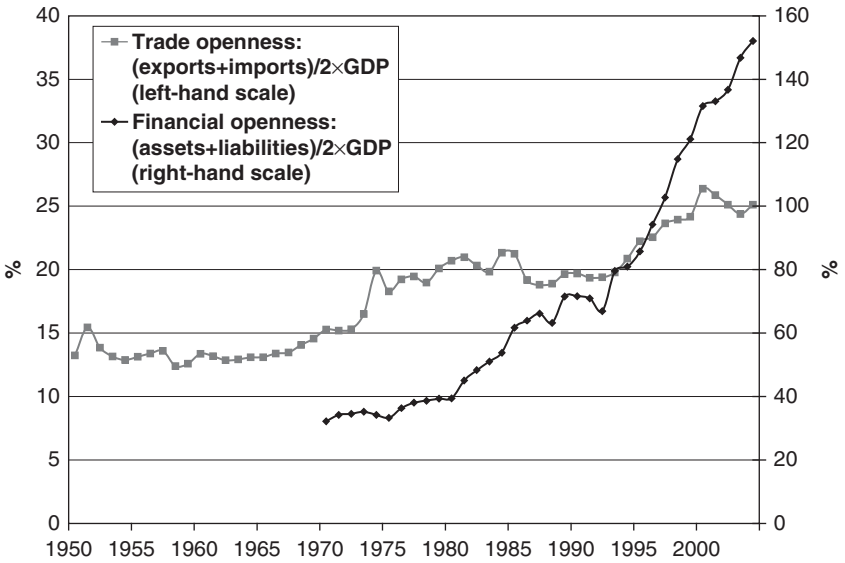


Figure 2.4 Trade and financial openness for G7 countries, 1950–2004 (unweighted averages).
Source: Authors' calculations based on Penn World Tables and Lane and Milesi-Ferretti databases.

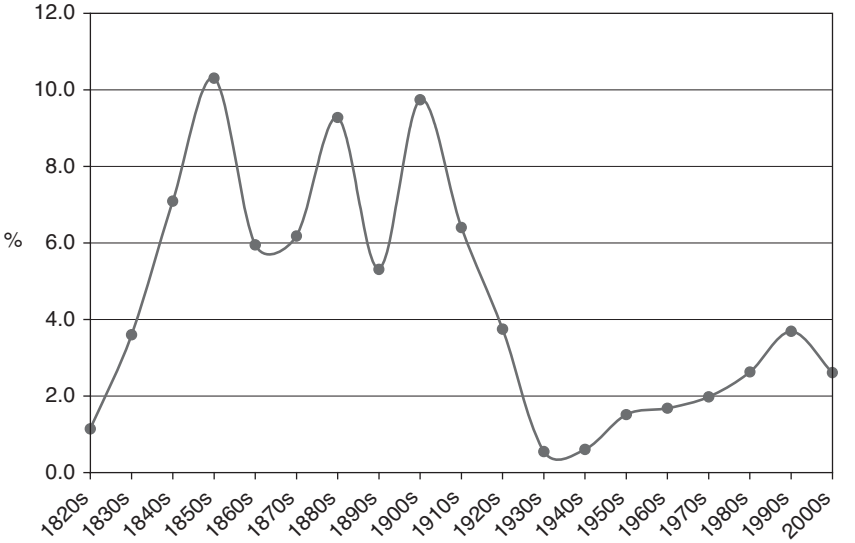


Figure 2.5 Inflows of migrants into the US as a percentage of the resident population, by decade.
Source: Authors' calculations based on US Census and US Homeland Security data.

asset and liability stocks is even more impressive, with their average proportion to GDP rising from 32% in 1970 to 152% in 2004, for G7 countries. An acceleration of this trend can be observed on the graph after the elimination of capital controls in the 1990s.

Ratios of this sort can be misleading, however, in suggesting that the world economy is close to being fully integrated. To start with, this is far from being the case for individuals. Even though international migration is on the rise, it does not compare to the massive flows observed in the nineteenth century. US data are telling in this respect: While inflows of new migrants amounted to 8% of the resident population in the 1850s to 1900s, the corresponding figure for the 1990s is only 4% (figure 2.5). The same holds for migration within the EU, which remains of limited magnitude in spite of the lifting of restrictions on movements of workers. Though some populations are mobile (for example football players, senior executives, and workers from the new member states), most European workers are not.

Furthermore, research has consistently pointed out the prevalence of a *border effect**: Product and capital markets are much less integrated across countries than within countries. This effect was first and amply documented for trade between the US and Canada, where, despite international integration through the North American Free Trade Area (NAFTA), McCallum (1995) found that trade between a pair of Canadian provinces was typically 22 times

Table 2.2
Theoretical and actual share of foreign securities in residents' portfolios, 2003

	Equity, %	Bonds, %
US: Actual	12.5	3.0
US: Theoretical	58.2	59.6
Germany: Actual	26.3	22.9
Germany: Theoretical	97.1	92.2
UK: Actual	45.7	69.4
UK: Theoretical	92.0	95.0

Note: For each country, the first row gives the actual share of foreign securities in equity and bond portfolio and the second the theoretical value of that share, based on the weight of the country in the world stock of the corresponding security. This simplified reasoning neglects the role of correlations of asset returns with each other and with household income, which may lead to another asset allocation.
Source: IMF, *World Economic Outlook 2005*.

greater than trade between a similar Canadian-province–neighboring-US-state pair. While qualifying these early findings, recent work[35] still points to a significant border effect. Such findings are not limited to trade between the US and Canada. In Norway, for example, only 39% of firms sell products on foreign markets and only 18% export more than 5% of their total turnover (Mayer and Ottaviano, 2007). Even in the EU, where all internal tariffs have been eliminated and all administrative obstacles to trade have been made illegal, two cities or regions trade 10 times more with each other when they belong to the same country than when they belong to different countries (Mayer and Zignago, 2005). Financial portfolios also remain biased toward domestic assets, whereas in an integrated economy portfolios would presumably include a fraction of the assets issued in each country (table 2.2).

It would therefore be confusing to start from the assumption that integration is perfect. It is intense, but far from total, and this complicates the allocation of policy responsibilities to the various possible levels of government.

2.2.2 International policy coordination

The management of economic interdependence between politically independent states was deemed sufficiently important to lead to the creation of several international bodies, from the International Monetary Fund in 1945 to the

35. Notably by Anderson and van Wincoop (2003), who develop a theoretically consistent model that explains part of the border effect and still find a ratio greater than 10, and Yi (2010), who develops an explanation based on multi-stage and sequential production and is able to explain away 3/8 of the empirically measured border effect.

*Organization for Economic Cooperation and Development** (OECD) in 1961 (it succeeded the OEEC formed in 1947 to administer American aid to Europe under the Marshall Plan), to the informal but powerful *Group of Seven** (G7) in 1975 (where it started as a G6), which later became the *Group of Eight** (G8, including Russia), and to the *Group of Twenty** (G20) in 1999 in the aftermath of the emerging market crises, whose wider and more diverse membership includes Argentina, Brazil, China, India, Indonesia, Russia, Saudi Arabia, South Africa, and Turkey.[36] As a collective international response to the global economic and financial crisis that started in 2007, the G20 was invited to take an increasing role in strengthening international cooperation. The final communiqué of the September 2009 Pittsburgh Summit recognized the G20 as the "premier forum" for international economic cooperation. The creation of those groups and of numerous other sectoral or ad-hoc groupings is an indication of the growing importance of *international economic policy coordination**.

Coordination has two main motives. The first is the provision of what has come to be called global public goods, such as the preservation of the global climate or international financial stability. The second is the optimization of policy outcomes when a country's decision significantly affects its neighbors.

a) Global public goods

To understand what a global public good is, it is useful to start from two important possible properties of the goods and services offered for consumption: Excludability and rivalry. Excludability means that consumption may be reserved to some individuals or households. Examples include standard private-consumption items such as clothing, food, and cars, but not clean air, free-to-air TV programs, security, and financial stability (those goods and services are deemed nonexcludable). Rivalry means that one's consumption reduces the availability of the good for others, which is true for fish in a lake but not for public lighting or street safety (those goods are deemed nonrival). The two properties are independent, which implies that all four cases can be found (table 2.3).

The social value of the consumption of a pure private good equals its private value (nobody but me cares about the comfort of my shoes). But this ceases being true for nonrival goods, because their consumption also has value for other people without reducing its value for me, which means that the incentive to produce them may be insufficient to ensure adequate supply. If I clean the street in front of my house, this has value for my neighbors too, and I am

36. The G7 was initially created as a G6 at the Rambouillet Summit in 1975 and included the US, Japan, Germany, the UK, France, and Italy. It was joined in 1976 by Canada. Russia is now also a member at the heads-of-government level (turning it into a G8). G8 summits are held once a year. Meanwhile, finance ministers continue to meet in G7 format, in particular around the spring and fall meetings of the IMF and World Bank.

Table 2.3
Excludability, rivalry, and the definition of public goods

	Excludable	Nonexcludable
Rival	Private good *Ex: Shoes*	Common good *Ex: A lake's fish resources*
Nonrival	Club good *Ex: Patentable inventions*	Public good *Ex: Financial stability*

therefore tempted to wait until they do so. Excludability can be a solution to the problem (through patenting an invention, one can acquire property rights and can charge others for the use of the invention[37]), but not all goods are excludable. For those that are neither rival nor excludable, there is no easy way to ensure adequate supply. Absent government intervention, the standard theory presented in chapter 1, therefore, suggests that the production of these goods will be sub-optimal.

This approach is used to frame the discussion on global governance. Climate preservation, sustainable management of depletable natural resources, and financial stability, to name but a few, are thus frequently deemed *global public goods**.

In practice, international cooperation in this field involves three major difficulties: First, how to agree on what constitutes an international public good? Second, what are the appropriate instruments or rules to produce it? Third, who should contribute to financing it? Global warming is a case in point where there have been disagreements at all three levels. The US administration under President George W. Bush for several years disputed the evidence of a link between global warming and carbon dioxide emissions (eventually recognizing it had relevance at the end of 2007). US policymakers and experts tend to put emphasis on research into new energy sources, clean technologies, and carbon sequestration and storage, rather than on binding, quantitative carbon dioxide emission targets as the Europeans do. And emerging countries such as China and India argue that they have not yet significantly contributed to the *stock* of greenhouse gases (which is true, despite the fact that they contribute significantly to the *flow*—as an example, China's emissions exceeded those of the US in 2007), and should therefore not be asked to curb their emissions at their current stage of economic development.

Similar issues arise for other potential global public goods such as international development. Failure to address mass poverty in poor countries would in the end reduce the welfare of citizens all around the world, either for reasons of altruism or because underdevelopment penalizes global prosperity

37. See chapter 6 for more on intellectual property.

and fuels terrorism, crime, the spread of diseases, and mass migration. But *overseas development assistance* (ODA)* also has a bilateral dimension between donors and receivers, either in view of geography and history, or because donors are using this lever to maintain some influence in specific regions. In effect, multilateral ODA represents only between 20 and 30% of total assistance. Also, the relative effort among donors is very uneven, with Sweden contributing more than 5.5 times the US or Japan in proportion to gross national income (1.12% against 0.20% and 0.18%, respectively, in 2009).

International cooperation is designed and enforced through various means. A particularly effective way is to agree once and for all on the rules of the game and enforce them. Once the rules have been defined and adopted, each player remains free of its decisions as long as they remain in conformity with the rules. International trade provides an example of such *rules-based cooperation* (or *coordination*)*: All the 153 members of the World Trade Organization have subscribed to the set of about 60 multilateral trade agreements covering goods, services, and intellectual property agreements and agree to abide by the WTO decisions in case of disputes.

A less-demanding form of cooperation relies on soft, rather than hard law. The coordination of bank supervisors also proceeds through common rules (the so-called Basel ratios), yet these have no legal value until they are enforced by national legislation. Another example is the promotion of standards and codes by the International Monetary Fund and the World Bank. The initiative covers policy transparency (e.g., data publication), financial sector regulation and supervision (e.g., banking supervision and regulation), and market integrity (e.g., corporate accounting and auditing) and it aims at promoting good practices through setting standards and reporting on each country's compliance with them (IMF and World Bank, 2005).

What this description suggests is that there is no single template for global governance, but rather a combination of various approaches and institutional set-ups.

b) International spillovers

The second motive for international policy coordination, and in fact the most traditional one, arises from *international spillovers** of economic policy. For instance, a rise in the US interest rate may lead asset prices to fall worldwide, emerging countries to default on their debts, or the dollar to appreciate against other key currencies. This is a typical case of an externality that will not be taken into account by the US government, unless there is some coordination with other countries, for instance within the G20.

This kind of spillover gives rise to strategic interactions between countries and implies that in the presence of significant cross-country spillovers,

separate decision-making by national governments may not be optimal. The pitfalls of such decisions and the essentials of coordination are well captured by the canonical "prisoner's dilemma" example (box 2.13).

Box 2.13 The "Prisoner's Dilemma" and the Shortcomings of Independent Policymaking

The *prisoner's dilemma** was first expounded at the *Rand Corporation* in 1950 (Tucker, 1950, 1980). It provides a simple example of strategic interdependence between separate decisions and illustrates the potential gains from coordination.

After a crime is committed, two suspects are jailed, awaiting judgment. Neither one acknowledges his own culpability. Absent compelling evidence, the judge establishes the following rule: If either one of the two suspects claims his innocence and denounces the other one, he or she will be released and the other suspect will be condemned to a fixed 10-year sentence; if the two suspects accuse each other, they will be considered jointly guilty and will be condemned, but their willingness to cooperate with the judge will be rewarded and they will be condemned to five years of prison only; finally, if both of them continue to assert their innocence and do not accuse each other, they will both spend one year in prison.

Each prisoner's fate thus depends on his or her own decision as well as on his or her fellow suspect's decision, which gives the problem the structure of a game. The square 2×2 matrix of table B2.13.1 gives the payoffs associated with the two prisoners' decisions in the form of (x, y) where x is the reward to prisoner 1 and y the reward to prisoner 2.

Table B2.13.1
The prisoner's dilemma

Prisoner 1	Prisoner 2	
	Betrayal of other	Cooperation with other
Betrayal of other	(–5, –5)	(0, –10)
Cooperation with other	(–10, 0)	(–1, –1)

To find out what is the optimal strategy for a given prisoner, say 1, the outcome of either decision must be examined depending on the other prisoner's decision. If prisoner 2 betrays his or her accomplice (first column), prisoner 1 will also find it beneficial to betray (because $-5 > -10$). If prisoner 2 cooperates with the other prisoner (second column), prisoner 1 will still find it beneficial to betray (because $0 > -1$). So in the absence of communication, it is individually preferable to denounce the

other prisoner. Both prisoners therefore get a five-year sentence. This is called the *noncooperative equilibrium**, or *Nash equilibrium**.

This equilibrium is not optimal: If the two prisoners were able to talk to each other and reach a deal, they would both be better off cooperating and remaining silent, in which case each of them would be sentenced to 1 year only (*cooperative equilibrium**).

This model shows that in the presence of interdependence, rational decentralized decision-making may not be optimal. It shows that cooperation can be beneficial, but suggests also that it can be difficult to reach and to sustain cooperation, because, once one of the players is convinced that the other will cooperate, he or she has an incentive to betray . . . and through subsequent retaliation both players return to the noncooperative equilibrium.

Because of its simplicity this model has been widely used as a reference for analyzing international policy coordination. However, it does not imply that formal cooperation is always necessary. It can be shown (for example, Axelrod, 1984) that such a game leads to a stable cooperative solution when it is played repeatedly over an infinite horizon with a simple retaliation rule: If one of the players cheats and does not cooperate, the other responds by not cooperating in the following round ("tit-for-tat" strategy).

More formally, the gain from coordination can be illustrated as follows. There are two symmetric countries,[38] home and foreign (the latter being denoted with an asterisk). As a consequence of interdependence, policy outcomes Y and Y^* not only depend on national policy decisions x and x^* but also on the neighbor's decisions. Hence,

$$Y = H(x, x^*) \qquad \text{(B2.13.1)}$$

and

$$Y^* = H^*(x^*, x) \qquad \text{(B2.13.1)}'$$

where Y, Y^* are n-dimensional vectors and x, x^* are scalars (but having more than one policy instrument per country would not affect the result). In each country, the policymaker aims at maximizing a social welfare function $U(Y)$. Because of (B2.13.1), $U(Y)$ can also be written $V(x, x^*)$. When acting in isolation, the national policymaker maximizes V, taking x^* as given. Thus, we have:

$$\underset{x}{Max}\, V(x, x^*) \qquad \text{(B2.13.2)}$$

and

$$\underset{x^*}{Max}\, V(x, x^*) \qquad \text{(B2.13.2)}'$$

38. The symmetry assumption is made for simplicity motives only. Removing it does not affect the result.

The first-order conditions give:

$$\frac{\partial V(x, x^*)}{\partial x} = 0 \qquad \text{(B2.13.3)}$$

and

$$\frac{\partial V^*(x^*, x)}{\partial x^*} \qquad \text{(B2.13.3)}'$$

This implies that each country's optimal policy depends on the policy of the neighbor. Formally, solving equations (B2.13.3) and (B2.13.3)′ yields two reaction functions:

$$x = F(x^*) \qquad \text{(B2.13.4)}$$

and

$$x^* = F^*(x) \qquad \text{(B2.13.4)}'$$

the intersection of which gives the noncooperative Nash equilibrium. It can easily be shown that this equilibrium is not a Pareto-optimum. The reason is that the Pareto-optimum is the solution of the following equation:

$$\underset{x,x^*}{Max}\, V(x, x^*) \quad \text{subject to} \quad V^*(x^*, x) \geq V_0^* \qquad \text{(B2.13.5)}$$

where V_0^* corresponds to a given level of utility for the foreign country. The corresponding Lagrangian is:

$$L = V(\boldsymbol{x}, \boldsymbol{x}^*) + \lambda[V^*(\boldsymbol{x}^*, \boldsymbol{x}) - V_0^*] \qquad \text{(B2.13.6)}$$

whose maximization implies:

$$\frac{\partial V}{\partial x} = -\lambda \frac{\partial V^*}{\partial x} \qquad \text{(B2.13.7)}$$

and

$$\frac{\partial V}{\partial x^*} = -\lambda \frac{\partial V^*}{\partial x^*} \qquad \text{(B2.13.8)}$$

This condition, which differs from (B2.13.2) and (B2.13.2)′, in fact corresponds to the maximization, not of V, but of $V + \lambda V^*$. In other words, independent policymaking does not yield an optimal result.

There are many examples of such interdependence giving rise to coordination problems. A simple one is that of two countries in a fixed-exchange-rate regime hit by a common shock that attempt to escape the adverse consequences of the shock on their external balance by running a restrictive fiscal policy (box 2.14). As the shock is a common one and both countries react in the same way, this attempt will in the end prove futile. The only effect

of running a more restrictive policy will be to lower output, not to improve the external balance.

Another example is monetary coordination under flexible exchange rates. Following a global shock, countries acting in isolation are inclined to use the exchange rate strategically: They may inflate excessively, depreciate their currency, and export unemployment following a recessionary shock or, on the other hand, run an excessively tight monetary policy, appreciate their currency, and export inflation following an inflationary shock. Of course, not all countries can export their unemployment or their inflation simultaneously because there are only $(n-1)$ exchange rates for n countries, so if all countries behave in the same way their exchange rates will remain unchanged and their attempt at exporting their difficulties will be frustrated.

There have been historical cases of such *beggar-thy-neighbor** policies: For example, the competitive depreciations of the 1930s, which contributed to the worsening of the economic and political climate in the aftermath of the Great Depression, or the US monetary policy of the early 1980s, which resulted in a sharp increase in interest rates (further compounded by the deterioration of the fiscal deficit induced by the Reagan administration's tax cuts and increases in military spending), large foreign capital inflows, and the export of inflationary pressures to the rest of the world through an appreciating dollar. Europe, also eager to combat inflation, had to embark on an ever more rigorous monetary policy. On the whole, the reduction of inflation had a growth opportunity cost higher than it might have been if the interaction had been taken into account and had led to a cooperative approach (which would have required a modification of US economic policy). A final example is that of East Asia in the 1990s, where countries were individually pegging their currencies to the US dollar instead of jointly adopting a basket reference. Lack of coordination prevented them from taking the decision that was in their common interest.

Box 2.14 A Bare-Bones Coordination Model

One of the simplest models of coordination is a two-country, symmetric model with an exogenous rest-of-the-world, under fixed exchange rates. Countries each have one single instrument, namely fiscal policy. Asterisked variables represent the foreign country, nonasterisked ones the home country. Fiscal expansion in each country has an effect on its neighbor, so that, if Y represents production (measured as the gap between actual production and full employment) and g the fiscal instrument:

$$Y = \phi g + \psi g^* - u \qquad \text{(B2.14.1)}$$

$$Y^* = \phi g^* + \psi g - u \qquad \text{(B2.14.2)}$$

where $\phi > \psi > 0$ and u represents a symmetric external shock (a variation in demand from the rest of the world). Let us suppose, further, that the

governments of both countries care about their external balance b, which is given by:

$$b = \rho(g^* - g) - u \quad \text{and} \quad b^* = \rho(g - g^*) - u \tag{B2.14.3}$$

where $\rho > 0$. If $u > 0$, i.e., if foreign demand falls exogenously, each country sees its income and its external balance deteriorate and would be interested in a fiscal expansion by its partner, which would boost its exports. Failing this, it will choose its fiscal policy in order to minimize a loss function: $L = \omega Y^2 + b^2$ and symmetrically $L^* = \omega^* Y^{*2} + b^{*2}$ where ω, $\omega^* > 0$ represent the home and foreign weights of income relative to the external balance in the loss function.

When countries act independently, the optimum policy for the home country is given by:

$$g = \frac{(\rho^2 - \omega\phi\psi)g^* + (\omega\phi - \rho)u}{\omega\phi^2 + \rho^2} \tag{B2.14.4}$$

A symmetrical result holds for g^*.

The reaction of each country to the exogenous fall in external demand ($u > 0$) depends on the relative weights of the internal and external objectives in its loss function (if $\omega < \rho\phi$, the government reacts to the shock by a fiscal contraction to restore external balance). But this reaction depends on the policy conducted by the other country: There is a *reaction function** that gives each country's optimal policy choice as a function of the other's.

However, the model being fully symmetrical, both countries will conduct the same policy, $g = g^*$ and therefore, at equilibrium:

$$g = g^* = \frac{\omega\phi - \rho}{\omega\phi(\phi + \psi)}u \tag{B2.14.5}$$

$$Y = Y^* = -\frac{\rho u}{\omega\phi} \tag{B2.14.6}$$

$$b = -u \tag{B2.14.7}$$

$$L = \left(1 + \frac{\rho^2}{\omega\phi^2}\right)u^2 \tag{B2.14.8}$$

The interest to coordinate arises from comparing the loss under the Nash equilibrium (equation (B2.14.8)) and the loss when both countries cooperate by jointly minimizing the sum of the two loss functions. In the latter case, they recognize that trying to cushion the impact of the external shock is useless; the optimal policy thus yields $Y = Y^* = 0$ and the loss is $L = L^* = u^2$. It can be seen that cooperation leads each government to be better-off.

This approach has a general scope. The recommendations to which it leads will have to be modified depending on the kind of external effects that are being considered. For example, if it is proved that the harmful effects of a deficit on the neighboring country (through, for example, the rise in the long-term interest rate) exceed its benefits, the purpose of coordination will be to limit the recourse to national deficits. However, coordination will still be necessary.

c) The limits and shortcomings of coordination

There are also strong arguments against coordination and some authors have shown that it could even sometimes prove harmful.

The first argument is that governments can cheat, and refrain from implementing agreed policies. This is an especially relevant concern when there is imperfect information about the other players' policies. The European soft coordination processes offer numerous examples of commitments that are not followed by actual policy decisions. Second, when there is uncertainty regarding the true model of the economy, errors in or disagreements among models can also lead to counterproductive coordination (Frankel and Rockett, 1988).[39] Third, coordination can be regarded as a form of collusion that prevents the emergence of adequate policies through a process of policy competition. Martin Feldstein (1988) argued early on that coordinating inadequate economic policies would lead to a result inferior to what could be achieved through noncoordinated, but better individual policies. He was in fact echoing fears regularly aired by Germany about either global or European coordination. Kenneth Rogoff (1984) has given a formal presentation of this argument in a setting where coordination weakens the commitment of central banks to noninflationary policies. Fourth, and not least, partial coordination can worsen rather than improve the policy outcome (this is a special case of the second-best argument discussed in chapter 1). For example, in a monetary union a coordination amongst budgetary authorities that does not involve the central bank can actually result in an inferior performance.

Few empirical studies have been conducted to evaluate the concrete benefits from economic policy coordination. The first and still most quoted one is that of Oudiz and Sachs (1984), where the authors evaluate governments' objective functions on the basis of observed past behavior and conclude that for the

39. Results are not robust against changes in specification. Ghosh and Masson (1994) show that uncertainties regarding the current economic situation, models, or the nature of external effects can actually strengthen the case for economic policy coordination and the benefits to be derived from coordination.

major countries, the gains from coordination are of the order of magnitude of half a percentage point of GDP. This is hardly enough to offset the limitations to national decision-making implied by coordination.

The lessons from experience also lead to mixed conclusions. At the global level, systematic coordination has never lasted long, but nonrecurring initiatives have taken place. Worth mentioning are the coordinated reflation engineered at the Bonn Summit in 1978, the *Plaza agreement** to amplify the depreciation of the dollar in October 1985, the January 1987 *Louvre agreements** to stabilize exchange rates, and the "Framework for strong, sustainable and balanced growth" initiated in 2009 by the G20 to address global current account imbalances. These examples do not provide an unambiguous demonstration of the benefits from coordination. The coordinated reflation carried out by Germany and Japan at the time of the Bonn Summit comes out as a failure *a posteriori*, notably because the economic context in which it was implemented (the second oil crisis) did not correspond to that in which it had been decided. It led to a resumption of inflation and to a current-account deterioration in Germany, thus feeding in that country a lasting mistrust of coordination. Regarding the exchange-rate agreements, the attempt to engineer coordination of monetary and budgetary policies through committing to exchange-rate levels was short-lived and it is not clear whether exchange-rate developments were actually attributable to the agreements. Furthermore, Japanese policymakers resent the episode because the constraints on monetary policy contributed to the mismanagement of the boom–bust cycle of the 1990s.

Coordination has been markedly more successful at the technical level, especially among central banks. Reactions to crisis episodes, be it the October 1987 stock market crash, 9/11, or the liquidity crisis of August–December 2007, were tightly and successfully coordinated between US and European monetary authorities.

Lastly, whereas the G7 and now the G20 play an undeniable role of exchange, information sharing and impulsion, which is indeed crucial for cooperation, they are often wrongly presented as major economic policy coordination bodies. In fact, as James Tobin (1990, p. 13) observed early on, the G7 "creates the shadow of coordination but not the substance." Monetary policy pertains to the competence of the independent central bank, while fiscal policy is decided by national parliaments. In this institutional context, the role of the G7 is more centered on information-sharing, communication toward financial markets, consensus-building, and the definition of common positions on global issues such as transition in the former USSR, emerging market crises, debt forgiveness for low-income countries, and financial stability. The G20 became a highly visible coordination platform for crisis management in 2007–09. On the whole, however, a dose of skepticism about the role of strategic economic policy coordination as a method of management of world interdependence is warranted.

d) Global institutions and governance

In all countries, economic policy is affected by the overlap of local, national, regional, and global institutions or agreements. The regional level, especially, has gained prominence since the early 1990s, as illustrated in figure 2.6. The most famous regional agreements are the European Union in Europe, APEC in the Pacific, NAFTA and Mercosur in America, WAEMU, ECOWAS and COMESA in Africa, and ASEAN in East Asia. However, according to the World Trade Organization (WTO) "nearly all of the WTO's Members have notified participation in one or more regional trade agreements (some Members are party to twenty or more)." Although free trade is generally the main objective of regional initiatives, other fields of economic policy also have regional dimensions, including monetary policy (through regional monetary unions), economic development (through regional development banks), or regional capital markets (e.g., the Asian bond initiative). The European Union provides the most far-reaching example of regional coordination in the world.

In contrast, the global dimension of policymaking has experienced limited progress since World War II. Even before the war ended, two global institutions—the International Monetary Fund and the World Bank—were created to accompany the re-opening of goods and capital markets, coordinate capital flows toward reconstruction and development, and provide financial assistance in cases of balance of payment difficulties. They were complemented by several rounds of negotiations under the General Agreement on Tariffs and Trade (GATT). The World Trade Organization (WTO) was created in 1995 with a dispute-settlement body that can be requested to resolve disputes between member countries. The creation of the WTO was concomitant to an expansion of participation in international trade organizations (153 countries

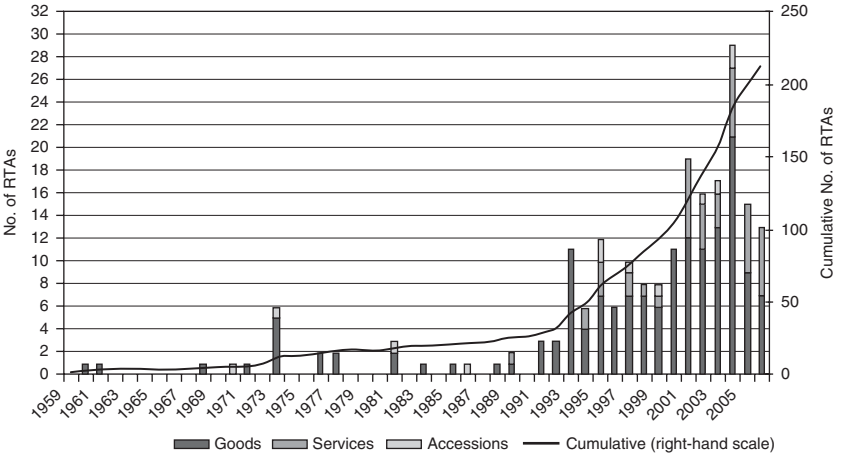


Figure 2.6 Regional trade agreements (RTAs) in force in 2010 by year of creation. Source: WTO. www.wto.org, April 2010.

were members of the WTO in 2010, compared to 23 countries participating in the GATT in 1947). Global trade negotiations have become more difficult, with the inclusion of large, developing countries, as well as with the emergence of sensitive negotiation topics such as agriculture, services, investment, and intellectual property.[40] Meanwhile, the world was shaken by a series of large financial crises in the 1990s and early 2000s, and a number of emerging countries (especially in Asia and Latin America) decided that they would be better off not relying on the IMF any longer. International instances of financial regulation, such as the IMF and the Basel Committee failed to predict and to prevent the 2007–08 financial debacle (whose implications, however, restored some reliance on and some faith about the IMF). Finally, despite increasing concerns in the scientific community about global warming, no institution has been created to deal with this issue. On the global level, only the United Nations has addressed the problem,[41] but the largest polluter, the US, has signed but not ratified the Kyoto protocol, which commits industrial countries to meeting quantitative targets for the 2008–12 period for their emissions of greenhouse gases. The December 2009 UN Copenhagen climate summit failed to reach a broad international agreement on country-specific, binding quantitative greenhouse gases emissions targets for the post-2012 period.

The expression *world* or *global governance**, which has gradually emerged as a widely used concept, refers to a problem on which many debates now focus: How to govern (a globalized world) without a (world) government? Or, in other words, how to fulfill, through a number of partial institutions and rules, functions which, within nation-states, usually fall within the prerogative of governments?

Although the institutions of global governance and the commonly agreed rules can be interpreted, like the process of European integration, as constraints on states' decisions, they are based on very different principles. The EU rests on an economic organization, but it is a political construct, with two consequences: Its competences are not limited *a priori*; the treaties which govern it are ratified (either directly or through their parliaments) by the people of the member states; and the laws that it enacts are jointly approved by the Council (which represent member states) and the European Parliament (which is directly elected). In contrast, governance by international institutions borrows more from the independent authority model. Admittedly, their decisions require the approval of the majority of the member states (in the case of the IMF or World Bank) or are taken pursuant to texts approved by them unanimously (the WTO), but each of them intervenes

40. At the time of writing this book, the "Doha Round" of multilateral trade negotiations, which opened in 2002, had not yet reached a successful conclusion. However, the WTO had played a useful role in containing protectionist pressures throughout the crisis.

41. Most notably and effectively through the Intergovernmental Panel on Climate Change (IPCC), a scientific body created by the World Meteorological organization (WMO) and the United Nations Environment Programme (UNEP).

in a specific area, delimited on the basis of an explicit mandate (cf. table 2.4). Their legitimacy derives less from the method of decision than from their specialization and from the manner, satisfactory or not, in which they fulfill their mandate (in other words, it rests more on their performance than on the quality of the decision-making process).

This system of governance is probably the only one possible, since there is no global political authority that has legitimacy to make and enforce choices: In the terms of chapter 1, there is no such thing a world social welfare function in the absence of world representation. The only *legitimate* deliberative body at the world level is the United Nations General Assembly, but the large number and the very heterogeneous sizes of the states which are represented make it quite ineffective for addressing global economic issues. World governance therefore involves several specialized multilateral organizations.

Beyond the fact that the policy conducted by a given international organization is a matter for debate, the de facto hierarchy of the existing international institutions can also be criticized for not reflecting the perceived priority of international problems. The power of the IMF and the World Bank, and the fact that WTO panels contribute to making international trade law when settling disputes between member states, are signs of organizational effectiveness. However, these very strengths also underline the relative inefficiency of the world governance system as regards environment, public health, or social and labor legislation. These areas lack specialized organizations of comparable effectiveness. The institutions dealing with finance and trade are not matched by organizations with competences and resources adequate to cope with, for example, climate change, major pandemics, or international migration. In a better-balanced institutional framework, these institutions could systematically be consulted when financial or trade decisions involve environmental, health, or social issues (Jacquet, Pisani-Ferry and Tubiana, 2002).

2.2.3 Federalism

The French Republic is meant to be "one and indivisible," and the F-word, *Federalism*, is banned from the UK political vocabulary, but there are close to 30 countries representing 40% of the world population that are officially classified as federations (or confederations).[42] Those countries are generally large ones such as Brazil, Australia, Germany, India, and the US, but Switzerland is a confederation too. In those countries a large proportion of policy decisions are taken at the state level.

42. The distinction between a federation and a confederation is not a clearcut one. Confederations are generally established by treaties and are more decentralized. The ultimate power rests with the participant states, whose unanimity is required for important decisions. According to such a distinction, the Swiss Confederation is, in fact, a federation.

Table 2.4
Scope, rules, and means of the major international organizations

Sector (institution, creation)	Voting rules	Institutional strength	Legal means	Financial strength
Trade (GATT, 1947 + WTO, 1995)	One country, one vote, simple or qualified majority for the application of the treaties, in practice consensus	Weak, except for dispute settlement	Arbitration and dispute settlement (through the Dispute Settlement Body)	Irrelevant
Currency and financial stability (IMF, 1945 + BIS, 1930 and FSB, 1999)	IMF: Constituencies with weighted voting rights, simple or qualified majority; in practice consensus BIS: Weighted voting rights FSB: Consensus	IMF: Strong institutional coherence plus strong G20 support BIS: Important via the central banks FSB: Strong G20 support	IMF: Limited power to set standards, indirect power on countries under IMF assistance BIS and FSB: Indirect standard-setting power	Major vis-à-vis countries requesting assistance (mostly poor countries), nil vis-à-vis surplus countries Potentially important via the central banks
Development finance (World Bank, 1945)	Like the IMF with greater role for developing countries	Same as IMF	Almost absent	Declining before the 2007–09 crisis as countries had gained access to financial markets, significantly expanding in the aftermath of the crisis
Environment (UNEP, 1972)	In theory geographical constituencies, in practice depends on the United Nations	Weak and dispersed	Weak	Weak
Health (WHO, 1946)	General Assembly: One country, one vote; Board: One person, one vote	Significant, but strong decentralization	Important (immediately enforceable health standards)	Limited
Labor (ILO, 1919)	Parity between governments, employers and employees. General Assembly: One country, one vote Board: Permanent seats for large countries	Weak	Weak (implementation of agreed standards left to the goodwill of member states)	Weak

Source: Jacquet et al. (2002).

Even in unitary states, some decisions are decentralized at the local government level, for instance at the regional or communal level. As shown in figure 3.2 in chapter 3, the degree of decentralization varies greatly from country to country. For instance, in Denmark and in Greece, neither of which is a federal state, the share of local expenditures in total public spending is respectively 38% and 5%. In fact, the transfer of important powers to sub-national entities has much progressed since the 1980s in countries like Belgium, Spain, and Italy. Even the UK (with devolution to Scotland of legislative and executive competences in 1998) and France (through small steps, but in an unambiguous direction since the 1982 decentralization law) are taking part in this movement, which also extends to many developing countries.

In addition, whether they like it or not, and although the EU budget is limited to about 1% of GDP, European member states are engaged in what specialists call "intergovernmental federalism" and what the former President of the European Commission, Jacques Delors, called a "federation of Nation States." In the EU, some economic policy competencies, such as trade policy, competition policy, or, for euro area members, monetary policy, have been devolved to the EU. Similar, though less-ambitious attempts at building international unions have been launched in most regions of the world.

For a large proportion of the world population, the reality is therefore one of multi-level government. In such a context, debates about who has competence are often more lively than those over substance. On tax or social policies, but also on fiscal policy or, in the UK at least, on the currency, proponents and opponents of a European policy are involved in constant disputes. The polemics between Catalonia and the Spanish State or between Quebec and the Canadian federation are equally harsh.

The economic theory of federalism makes it possible to clarify these discussions and to propose criteria regarding the distribution of economic policy competences within a federation or guiding participation in an international union.

a) The economics of federations and international unions: A primer

The theory of "fiscal federalism" is limited neither to the study of federations—it deals in general with the "vertical" distribution of competencies between administrative and political entities—nor to fiscal matters—all policies are in fact concerned. It aims first at determining the level at which it is relevant to make particular decisions.

The basic rule is fiscal equivalence (Olson, 1969), which establishes that the administrative and financing organization of a public policy should coincide with its geographical impact area. The idea goes back in fact to Adam Smith, for whom:

> Those local or provincial expenses of which the benefit is local or provincial (what is laid out, for example, upon the police of a particular town or district) ought to be defrayed by a local or provincial revenue, and ought to be no burden upon the general revenue of the society. It is unjust that the whole society should contribute toward an expense of which the benefit is confined to a part of the society.
>
> A. Smith (1776, book 5, chapter 1, conclusion)

In more modern terms, the distribution of competences should be tailored to eliminate positive externalities deriving from one locality providing a good (such as an entertainment infrastructure) that also benefits other localities. The same applies to "internalities," i.e., cases where the impact of a policy is concentrated on a smaller area than that supporting its administration and financing (for instance, a local transportation system subsidized at the national level).

This matching rule between those who pay and those who benefit from public spending does not by itself involve either a centralization or a decentralization bias. It implies on the contrary that excessive centralization is as ineffective as excessive decentralization, and justifies the coexistence of several (possibly overlapping) levels of government in order to adapt as much as possible the management of the policies to the spatial distribution of their effects (for example, river pollution should be addressed by all localities on that river, whatever their country, and only by them). This rule is far from being respected in practice: Very often, local governments raise taxes to finance infrastructure or services (sport or cultural facilities for example) that may benefit residents of neighboring localities. By the same token, expenditures at the national level may benefit the residents of neighboring countries: For example, the road transport infrastructure in France benefits Northern European citizens when they travel to reach Southern European beach resorts.

The limit of such an approach obviously resides in the resulting complexity. However, it sets an important principle, with practical applications such as unions between neighboring localities and *enhanced cooperation** between neighboring states, involving, for example, within the EU cross-border cooperation between residents from the Mediterranean or from the Baltic.

Regarding the choice between centralization and decentralization, economic motives, which are only part of the decision criteria, give preference to decentralization. The Oates (1972) decentralization theorem claims that, in the absence of externalities and of economies of scale, decentralization is always preferable, or at least equivalent, to centralization. This is because different localities will weigh differently public goods against private ones in their consumption baskets. Hence, local decision will perform better in meeting taxpayers' preferences. To the extent that the public good provided in one locality has no impact on welfare in neighboring ones (no externalities),

and that the public good is not provided more efficiently at a centralized level (no economies of scale), this implies that provision at the most decentralized level is preferable.

This result holds, however, only in the absence of externalities and economies of scale. For certain goods centralization involves benefits, either because the benefit from their production cannot be restricted to the residents of one particular country, or because it involves increasing returns. This is, for example, the case for defense, research, and environmental policies.

When there are both heterogeneous preferences and externalities across jurisdictions, decentralization permits sticking with local preferences, but centralization allows internalizing externalities. There is thus a trade-off, and the optimal level of decision depends both on the size of the externality and on the divergence in preferences. This trade-off is at the core of the theory of international unions formalized by Alesina et al. (2005). In their model (box 2.15), centralization provides efficiency gains due to economies of scale or positive externalities between countries (or regions); decentralization allows each government to stick to national preferences between public and private goods. They show that entities characterized by heterogeneous preferences, for example, on the nature and on the volume of production of public goods, can nevertheless gain from collective action, because the welfare benefits from higher efficiency outweigh the welfare costs from the loss of autonomy. Again, national defense provides a simple illustration of their point: Few countries can afford a capacity to project their military forces beyond their borders. However, if they join an international cooperative body on foreign security, they can benefit from the expenditures of each member of the union, which they have to balance against the cost of having to agree with partners on the priorities and the practicalities of defense policy.

The same point can be made with negative externalities. For instance, suppose that the citizens of country A have a high aversion to income inequalities, whereas those of country B have a low aversion to it. A higher level of redistribution in country A will follow. However, this may induce migration of low-income persons from B to A. Due to budget constraints, country A may have to limit redistribution. Such a "race to the bottom," which is discussed in chapter 7, may lead governments to depart from their citizens' preferences, hence reducing the benefit of decentralization. In the EU, there is an ongoing debate on the issue of tax harmonization. Its proponents argue that, due to the mobility of skilled labor and capital, full decentralization leads to shifting the burden of taxation onto unskilled labor, despite the citizens' preferences. Its opponents claim that tax cooperation is an infringement on national sovereignty and will ultimately lead to a transfer of the taxing power to a (in their eyes illegitimate) central (supranational) authority.

Box 2.15 A Theory of International Unions: The Model of Alesina et al. (2005)

Alesina et al. (2005) formalize the creation of international unions as a response to a trade-off between preference heterogeneity and positive externalities.

A union is made of N countries assumed to be similar in size, which liaise to cooperate in the provision of a public good. Each country can provide the public good independently, but then it does not benefit from positive externalities deriving from other countries also providing the public good. Hence, the good considered is not a "pure" public good whose provision benefits all countries, but a "club" good for which exclusion is possible. Defense, or a common energy infrastructure, are examples of such public goods.

The cost of participating in the union comes from the fact that the N countries have different relative preferences for private versus public goods: Some prefer a lower quantity of public goods and a higher disposable income for private consumption (possible divergent preferences among various public goods of different nature are ignored). However, while taking part in a union, they opt for a collective choice that will be determined by a vote. They therefore face a trade-off which determines their participation in the union.

The utility function of the representative agent of country i ($i = 1$ to N) is specified as:

$$U_i = Y_i - G_i + \alpha_i H(G_i) \tag{B2.15.1}$$

if the country does not take part in the union. If it takes part in the union:

$$U_i = Y_i - G_i + \alpha_i H\left(G_i + \beta \sum_{j \neq i} G_j\right) \tag{B2.15.2}$$

where Y_i is income in country i, G_i the level of public good provision by country i, financed by taxation so that $Y_i - G_i$ is the disposable income for private consumption, and ΣGj is the public good provision by all other countries of the union. α_i measures for each country i the relative preference for consuming the public good rather than the private one. β (which lies between 0 and 1) measures the externality of other members of the union providing the same public good, and H is an increasing and concave function ($H' > 0$, $H'' < 0$), meaning that more public good provides higher utility to the representative agent, but with a decreasing marginal utility.

When each country chooses its level of production of the public good independently, it neglects the positive externality that this production could have on its partners, and therefore chooses a production level below the optimum. This is a typical case of coordination failure. When the N countries take part in the economic union, the level of public good provision is decided by a simple majority. It therefore corresponds to the preferences of the median country (for which the preference α_m lies in the middle of the distribution). The resolution of the model yields:

$$\alpha_m H'\left[\tilde{G}_N(1+\beta(N-1))\right] = \frac{1}{1+\beta(N-1)} \qquad \text{(B2.15.3)}$$

which defines the level of optimum public good production $\tilde{G}_N$ as an increasing function of:

- the number of members N,
- the strength of the externality β, and
- the preference α_m of the median country for the public good.

This is a standard application of the median voter model (box 2.10): What matters is not the average preference, but the preference of the voter who "makes" the decision. Obviously, this result depends on the voting procedure (simple majority, qualified majority, unanimity). In the event of unanimity (required in the EU for amendments to the treaty and for budgetary decisions), it is the preference of the country least favorable to the public goods which determines the decision (an example is the role of the UK in framing minimum labor laws at the EU level).

Let us now suppose that the union is already made up of countries 1 to $M(M < N)$ with contiguous preferences, so that (without loss of generality) for $1 \leq \ldots k \ldots \leq M$, $\alpha_1 \ldots \leq \alpha_k \ldots \leq \alpha_M$, and for $j > M$, $\alpha_j \notin [\alpha_1, \alpha_M]$). Upon entering the union, a new member will benefit from the external effects from which it was excluded before. For those that are already members of the union, the arrival of a new member will also cause a positive externality since the newcomer will spend $\tilde{G}_{M+1}$, but it will also modify the domestic equilibrium by moving the median voter.

Three interesting phenomena appear:

- The inclusion of an additional member whose preference for the public asset is low (α small) may nevertheless lead to an increase in the production level $\tilde{G}$, if the external effect overrides the displacement of the internal political equilibrium. Indeed, for a given median voter and due to positive external effects, the public good production level is an increasing function of the number of members.
- A majority of the former members may lose through the process of enlargement.

- If M countries with contiguous preferences have already formed a union, the $N - M$ remaining countries may not find it beneficial to join the union, because they would lose more from adopting a level of public good production quite different from their own preferences than they would gain from the benefits derived from other countries' expenditures.

The model can also be used to study the effect of enhanced cooperation in areas where national and common policies appear complementary.

This representation provides a useful framework for thinking about concrete issues. One example is the difference in attitudes toward the Kyoto protocol on greenhouse gas emissions. The US government under President George W. Bush rejected the protocol while the European governments endorsed it. Beyond politics, this difference in attitude signals a heterogeneity of preferences that can be interpreted as resulting from different patterns of urbanization and transportation, but also from a divergence between the US, where population density is low, winter harsh, and summer hot, and a densely populated Europe where the weather is milder and energy consumption lower. In the same vein, the French, whose country is less densely populated than The Netherlands, are keener than the Dutch on maintaining economic activity in the rural areas and on ensuring that the postal service reaches every remote corner of their national territory.

A common political argument in favor of decentralization is that it provides a guarantee against the confiscatory power of the central government.[43] A central state that is strong enough to exercise its basic functions may also be strong enough to confiscate private wealth. According to this view only basic economic functions must therefore be assigned to the central level—primarily, the management of a single market—while policies likely to have marked distributive effects should be left to the decentralized governmental levels. Competition between decentralized jurisdictions will ensure that none of them will resort to confiscation. Federalism within nations or international integration between them therefore counterbalances the tendency of states to behave like Leviathans and acts as a remedy to the alleged deficiencies of democratic systems. Indeed, following a famous expression introduced by Tiebout (1956), citizens have the ability to "vote with their feet," which may force elected officers to respect the citizens' preferences, even when the officers' re-election is not directly threatened. Decentralization can therefore be preferred independently from any welfare benefit due to the existence of economies of scale, of external effects, or of preference heterogeneity.

43. See for example Weingast (1995).

In Europe, capital mobility is high, and this conception of decentralization as a protection against Leviathan-like governments is widely voiced. For example, one cannot understand the European debate between tax competition and harmonization without referring to it. To see merits in fiscal competition, governments must, for example, be assumed to indulge in predatory behavior and to tax capital beyond what is economically justified in order to finance public expenditure (in particular income transfers) likely to bring them votes.[44]

b) The European Union

The *European Union**[45] was founded as the *European Community** in 1957 by six countries (Belgium, France, Germany, Italy, Luxembourg, and the Netherlands), which had previously successfully experienced cooperation amongst former foes within the framework of a *European Coal and Steel Community**. As of 2007, it comprised 27 countries, including 10 from the former Soviet bloc. It was initially created as a mere customs union complemented by common policies in a few sectors, yet one equipped with a sophisticated institutional and legal system. Over time, the EU has gradually gained competences over a wide range of policy areas and moved to a single market (see box 2.16). In the early 2000s, an attempt was made to equip it with a constitution, but the corresponding agreement was rejected in popular referendums in France and The Netherlands. Nevertheless, most of the provisions of the still-born constitution were in 2007 made part of a new treaty, the Treaty of Lisbon, which entered into force on 1 December 2009 after ratification by all 27 member states.

Box 2.16 Various Types of Economic Unions

In a *free trade area**, goods manufactured in the participating countries circulate duty-free, but each state keeps control of its trade policy with third countries. For example, the US, Canada, and Mexico, associated since 1992 within the North-American free trade Agreement (NAFTA), do not apply the same customs duties on imports from Europe. The management of a free trade area is complex, because these tariff divergences create an

44. For a discussion, see Tabellini and Wyplosz (2004). For a more general discussion on competition between states in the European context, see Pisani–Ferry (2004).

45. The EU was initially called European Community. It was renamed European Union in 1993 when the Treaty of Maastricht entered into force and added political and foreign policy dimensions to the initial economic dimensions of the Community. Here we refer to the *Union*, and keep *Community* only when referring to the past.

incentive to fraud. *Rules of origin** must be established that specify the conditions for a product to be regarded as actually originating in a country of the area.

In a *customs union**, all the imports of the union from the rest of the world are affected by the same customs duties, whatever their places of entry and of destination. For example, a Korean television set bears the same duties, whether it is imported by Antwerp or Barcelona, or sold in Prague or in Rome. The management of a customs union is simpler, but it requires that the participating countries adopt a common external tariff, as the Europeans did with the Treaty of Rome of 1957. However, a customs union does not necessitate the removal of border controls: Up to the 1992 single market, intra-European imports were subject to customs control, in order to check that they were in conformity with the national legislation. For example, and in order to slow down imports of video tape recorders, France was able, in the early 1980s, to temporarily impose systematic customs clearance in a single provincial city.

A *single market** is more ambitious: It requires removing obstacles to the mobility of goods and to the freedom to provide services, and extending mobility to workers and to capital. That requires the harmonization of regulations, in particular on technical or health standards, whose disparity would hamper the mobility of goods without border controls, and therefore the adoption of common regulations or the mutual recognition of national regulations. Within the EU, the single market rules have been in force since 1992, except that the 2004 and 2007 enlargement countries have been given transitional periods. Free provision of services is, however, hard to enforce. In 2005, a draft directive[46] aimed at organizing the services market through generalizing the country-of-origin principle (meaning that an accountant whose professional qualifications had been recognized in any of the EU countries could provide services to clients in the 26 other countries without further procedures) was rejected. Instead, the provision enacted gives states the right to legislate on the provision of services, provided it is done in a nondiscriminatory way.

A *monetary union** requires the adoption of a single monetary policy and therefore of a common central bank. Following an exchange-rate cooperation mechanism set up in 1979 (the European monetary System), the Economic and Monetary Union was negotiated in 1991, and after a transitional period, 11 countries adopted the euro on 1 January 1999

46. Directives are pieces of legislation that are adopted at EU level and thereafter transposed into national legislations. They oblige the member states to achieve a certain result but leave them free to choose how to do so.

(all EU members except the UK, Denmark, Greece, and Sweden). Greece joined in 2001, Slovenia in 2007, Cyprus and Malta in 2008, Slovakia in 2009. For most of the larger new member states, however, membership in the euro remains a distant prospect.

The *Economic and Monetary Union (EMU)** refers to all the provisions of the treaty that concern the single market, fiscal policies, and the single currency.

The preference for decentralization is reflected in the EU by the subsidiarity principle according to which policies should be assigned to the lowest level of government except when centralization is justified by the need to conduct joint action.

However, the subsidiarity principle has hardly resulted in a clarification of responsibilities (box 2.17). Since the 1990s, new forms of intergovernmental cooperation have developed, and coordination procedures have been strengthened that represent soft constraints on the member states' autonomy. A new method of governance based on voluntary intergovernmental cooperation called the *Open Method of Coordination** has even emerged and is being used for coordinating policies in fields like labor markets and research, competence for which remains in the hands of the member states.[47]

Most questions raised by the theory of fiscal federalism are relevant for the study of European integration:

- Even if the Union is a political construct, economic efficiency arguments generally carry a greater weight within the Union than they do within individual member states in deciding what the Union should do;
- Even though the diversity of preferences has clearly shrunk in some areas (for example, price stability), it remains patent and it has increased with enlargement. This calls either for decentralization in areas (such as social areas) where these preferences differ, or for *enhanced cooperation** between states which exhibit similar preferences.[48]

47. This process is usually referred to as the Lisbon process since the initial decision to launch it was taken in 2000 at a heads-of-state meeting in Lisbon, Portugal.

48. Enhanced cooperation enables a sub-group of at least nine member states to cooperate within the framework of the Union's nonexclusive competences and to this end make use of the Union's institutions. Such cooperation must be open at any time to all member states and all can participate in corresponding deliberations (but not vote). Examples may include cooperation at the regional level or on matters on which not all members agree. However, this provision, initially introduced in the Maastricht Treaty in a slightly different form, has never been used. See Coeuré and Pisani-Ferry (2005) for a discussion of the concept applied to economic policies and to the Euro.

Box 2.17 European Union Principles

The European Community did not result from an economic project, but from a political ambition. The choice to start with coal, steel, and agriculture resulted from the will "to create, by establishing an economic community, the basis for a broader and deeper community among peoples long divided by bloody conflicts; and to lay the foundations for institutions which will give direction to a destiny henceforward shared" (from the European Coal and Steel Community Treaty of 1951). The integration method adopted was intended to pave the way for further, tighter integration: Rather than laying down the objective and criteria in advance, all the Community mechanics aimed at creating a dynamics of cumulative integration. According to the Treaty:

- The process is one of "creating an ever closer union among the peoples of Europe" (Preamble of the EU Treaty);
- The mechanism *of the acquis communautaire** makes the transfer of a competence to the higher, central level irreversible and binding for all new members;
- The Union can gain new powers if needed "to attain one of the objectives set out in the treaties." If "the treaties have not provided the necessary powers, the Council, acting unanimously on a proposal from the Commission and after obtaining the consent of the European Parliament, shall adopt the appropriate measures" (Article 352 of the treaty on the Functioning of the European Union).

Whereas a constitutional approach would have determined *ab initio* the distribution of competences, the European method therefore relies on a combination between small steps pragmatism and powerful lock-in of mechanisms integration.

The Maastricht treaty of 1993 started to challenge this logic with the introduction of the subsidiarity principle whose aim was to avoid excessive centralization. The EU treaty includes no less than three principles whose aim is to put a check on excessive centralization: The *principle of conferral* states that "the Union shall act only within the limits of the competences conferred upon it by the member states in the treaties to attain the objectives set out therein," (article 5-2) the *principle of subsidiarity** states that "in areas which do not fall within its exclusive competence, the Union shall act only if and insofar as the objectives of the proposed action cannot be sufficiently achieved by the member states," (article 5-3) and the *principle of proportionality* states that "the content and form of Union action shall not exceed what is necessary to achieve the objectives of the Treaties" (article 5-4). The introduction of those principles in the Treaty indicates that at a time when additional competences were conferred on

the Union, member states were concerned about the risk of excessive centralization.

In order to justify transferring a competence to the EU, therefore, it is no longer enough to prove that decentralization is not optimal, but (which is more demanding), one must demonstrate that centralization is necessary. This distinctly differs from the principles that presided over the creation of the Community. Indeed, the first European common policy initiative, the European Coal and Steel Community of 1951, relied on a mixture between political and efficiency motivations. Its founders did not feel compelled to prove that centralization was necessary to achieve a free market for coal and steel. Later on, the creation of the common agricultural policy did not have to pass the test of whether centralization would perform better than decentralization either.

Five categories of competencies of the EU stand out in the current governance structure of the Union:[49]

- "Exclusive" Union competences. Here, "only the Union may legislate" and states can adopt legislation "only if so empowered by the Union or for the implementation of Union acts." This involves primarily trade, competition, fisheries policy, and, for euro area members, monetary policy.
- "Shared" competences for which *the initiative* belongs to the Union. States "shall exercise their competence to the extent that the Union has not exercised its competence". However, they can act insofar as the Union did not exert its rights or decided to cease exerting it. This primarily involves the management of the internal market, regional environmental policies, the common agricultural policy, consumer protection, transport, and energy. Moreover, the Union and member states can act jointly as regards research and development and humanitarian aid.
- The coordination of the economic and employment policies of the member states.
- The definition and implementation of a Common Foreign and Security Policy.
- Competences to support, coordinate, or supplement the action of the member states in specified areas such as health, industry, culture, or education.

Two features stand out in this list: First, the truly federal character of the Union, since its competences are in some areas higher than those of the member states; second, the complexity of the European decision system, since five categories of competences coexist that are not always easy to distinguish

49. As listed in Art. 2 to 6 of the Lisbon treaty.

Table 2.5
A simplified outline of competence assignment within the EU

	Member states	Union
Allocation		
Regulation of markets for goods and services[a]	X	XX
Regulation of capital markets	X	XX
Regulation of labor markets	XX	X
Infrastructures, research, education	XX	X
Farm support	—	XXX
Stabilization		
Monetary and exchange rate policy (Euro area)	—	XXX
Fiscal policies	XX	X
Redistribution		
Interpersonal (direct taxation, social transfers)	XXX	—
Interregional	XX	X
International (within the Union)	—	XXX

[a] Including indirect taxation
Key: By convention, the amount of the X is for each line equal to three. XX in a column indicates that the principal competence belongs at the corresponding level. XXX indicates exclusive competence.

and overlap each other, and since the logic of policy assignment across areas does not appear clearly.

From an economic (and nonlegal) perspective, it is possible, however, to give a simplified representation of the distribution of the major competences (cf. table 2.5).[50] Table 2.5 is schematic, but it has the advantage of revealing the economic logic of European integration, which rests on a number of hypotheses:

1. *Goods and capital are mobile between the countries of the Union, but labor is almost not.* This justifies assigning to the Union the regulation of the markets in goods, services and capital, and maintaining the member states' primary responsibility for labor market regulation. This very imperfect labor mobility is against the declared ambitions of the European single market (Article 45 of the Treaty states that within the Union, "freedom of movement for workers shall be secured") but it remains a fact, and is probably only partially explained by linguistic barriers; regulatory heterogeneities, for example, as regards pensions, contribute to it and member states hardly make efforts to reduce them.[51] Low labor mobility moreover enables them to protect their responsibility for redistribution policies between individuals, because

50. This presentation draws on insights shared by Tommaso Padoa Schioppa.
51. At the time of the negotiations on enlargement, the 15 former members of the Union moreover asked for, and obtained, a long transitional period (seven years) for the application of the freedom of movement for the nationals of the 10 new members.

the probability that differences regarding the degree or the methods of redistribution induce major population shifts between countries is low.[52]

2. *The management of the single market is a Union responsibility but states remain in competition for other allocation policies.* The exclusive or principal competences of the Union largely derive from the management of the single market, as is the case with international trade (transferred to the Union from the creation of the customs union, see box 2.17) or with competition, an area where the European Commission was assigned the double role of controlling concentrations and of policing state aid to companies. On the other hand, member states keep the principal responsibility for the other allocation policies, in particular for those (infrastructures, research and innovation, education) that are decisive for long-term growth. In those areas, the Union primarily plays a supporting role, via its budget.
3. *The single market calls for a single currency.* While the monetary union project partly fulfilled political aims, its economic justification was that the collective benefits from a single market where goods, services and capital would circulate without obstacles could be achieved fully only by ensuring exchange-rate stability, which is a major determinant of relative price stability (cf. chapter 5). However, under a regime of free capital mobility, exchange-rate stability could not be achieved in any systematic manner if countries maintained separate monetary policies. Hence, monetary union was necessary. The impact of monetary union on trade is lively debated in the academic community. In a famous paper, Andrew Rose (2000) predicted that monetary union might as much as triple intra-union trade. This evaluation has subsequently been much discussed and downplayed, but a common currency is generally viewed as having a positive impact on trade due to lower transaction costs and risk, higher transparency, and lower entry costs in foreign markets.[53]
4. *The single currency does not imply a federal budget but calls for joint surveillance of national fiscal policies.* This point has been the subject of very lively debates, particularly in connection with the *Stability and Growth Pact*, which provides a framework for national fiscal policies to avoid "excessive deficits." In the 1970s, there were some proposals to adopt a federal budget amounting to about 5% of GDP (McDougall, 1977). In fact, the adoption of the euro eventually took place *without any increase* in the Community budget. With a ceiling of 1.045% of

52. This, however, is debated. The German economist Hans-Werner Sinn (Sinn and Ochel, 2003) thus advises that migrants get the same benefits as residents only after they have been employed for some time in the host country. In between, they could remain eligible to social benefits from their country of origin.
53. See Baldwin (2006).

GDP for the 2007–13 period, this budget represents one fortieth of total public expenditure in the Union (figure 2.7 and box 3.15), and it is therefore excluded from playing any significant macroeconomic role. Such a role pertains to the common monetary policy and to national fiscal policies, in an individual or coordinated way. At the same time, however, fears exist that the lack of fiscal discipline in member states could challenge monetary stability, the central bank's mission. This is the justification for the Stability and Growth Pact (whose features are presented in chapter 3). At the end of the day, the stabilization function appears poorly defined in the Union: It partly pertains to the central bank in charge of monetary policy (but the central bank was assigned price stability as its overarching priority), and partly to national budgets (but they have to abide by the provisions of the Stability and Growth Pact). This indetermination leads to tensions, which are discussed in chapters 3 and 4.

5.

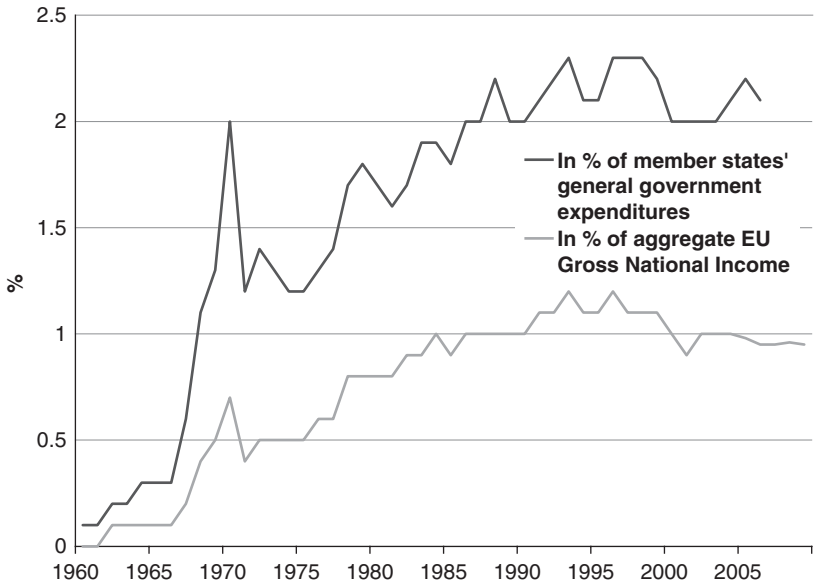


Figure 2.7 The budget of the European Union, 1960–2009.
Source: European Commission.

6. *The Union does not intervene in interpersonal redistribution, but fulfils a role of interregional and international redistribution.* In its early years, the Union had almost no redistributive role, except toward farmers through the Common Agricultural Policy (CAP). Spending on regional development represented just 3% of the EU budget in 1970 and 10% in 1980. In the 1980s, enlargement to include less-developed

countries (at the time, Spain, Portugal, Greece, Ireland) and the fear that the single market could lead to a concentration of economic activities in the most prosperous regions led to a significant development of regional policies: *Structural Funds**, which finance investments in less-developed regions and countries, amounted to 30% of total EU spending in 2000–06 (see chapter 6). In the 2000s, however, enlargement to Central and Eastern Europe did not give rise to a further increase in structural spending in spite of the significant income gap between the old and the new member countries. Redistribution from the former to the latter took place through redeployment of existing programs.

These five features summarize the current bases of EU governance—admittedly at the cost of ignoring some historically important and still-existing dimensions, such as the postwar goal of food independence that inspired the common agricultural policy.[54] They define a specific model of international union that has no equivalent in the world. While there are indeed several hundred regional unions (remember figure 2.4), the vast majority of them are limited to, at best, organizing a free trade area. One of the most advanced ones, the North-American Free Trade Area (NAFTA), which includes the US, Canada, and Mexico, does not include any supranational authority. Integration within the EU, however, falls short of the degree of integration prevailing in existing confederations or federations, such as Switzerland, the US, or Canada. Those are also based on a single market and a single currency, but they have in addition a common labor market and are endowed with much more important budgets (from 10% to 25% of GDP) that contribute to stabilization and interpersonal redistribution.

This European model took shape through a series of bold initiatives followed by crises and eventually political compromises, occasionally buttressed by economic analyses and recommendations. The creation of the Common Market in 1957, a combination of customs union and common sectoral policies, came after participating countries had failed in their attempt at creating a European Defense Community. The adoption of the single market was the political outcome of a compromise between the free-marketers, who saw in market integration a way to liberalize the markets in goods, services, and capital, and the federalists, who saw in the same liberalization a way to promote European integration.[55] Meanwhile, from an economic perspective, the single market responded to a powerful analytical insight that saw in the constitution

54. The CAP is increasingly criticized and discussed, the more so as the gradual shift from support for production to support for farmers' income has transformed it from a policy of allocation into one of redistribution to the benefit of a very specific category of the population.

55. Or, more bluntly, a compromise between UK Prime Minister Margaret Thatcher, a staunch advocate of liberalization and national sovereignty, and EU Commission president Jacques Delors, a former center-left French minister and an adamant federalist.

of a unified market a means of increasing welfare and of promoting growth (cf. chapter 6). Similarly, the single currency was the outcome of a political compromise between France, which saw it as a potential instrument of European power, and Germany, which could through the EMU export its own economic policy design: Between, therefore, François Mitterrand and Helmut Kohl. However, from an economic viewpoint, it largely borrowed from the theory of credibility and from theory-rooted arguments in favor of central bank independence (cf. chapter 4). However, the criteria elaborated for constituting the euro area largely ignored the lessons from the theory of optimum currency areas, which could have resulted in excluding some countries, or the lessons that could be learned from the US experience (cf. chapter 5). Finally, enlargement in 2004 and 2007, the (delayed) result of the dissolution of the Soviet bloc, was widely interpreted as a European reunification of historical significance. Yet, through the anchoring of their economic institutions and the resulting development of trade and foreign direct investment, the expected integration of the new member states played a major role in their successful economic transition.

Is the EU model now stable? The crash of the constitutional project, which was rejected by French and Dutch voters in 2005, is strong indication of the degree of doubt that remains among citizens. But the Treaty of Lisbon adopted in 2007 in place of the draft constitution retains many of its provisions and now provides the legal framework for EU policies (see footnote 10 in chapter 1). However, from its geographical scope to its governance and, more fundamentally, its role in the context of a globalizing world economy, many uncertainties remain, which have been put to the fore by the 2010 crisis in the Eurozone.

The future of the EU elicits as much debate among economists as it does among political scientists, politicians, and the public. Some (Alesina and Wacziarg, 1999) draw from the theory of federalism to claim that Europe has gone too far and intervenes in areas that should remain national prerogatives. Others, on the contrary, call for the current degree of integration to be deepened in order to move toward a "European Republic" where national economic policies would be subject to joint decisions (Collignon, 2002), or advocate in a more pragmatic way reforms likely to improve European governance (Sapir et al., 2004).

Contrary to what the theory of federalism suggests, the European system is characterized by a significant overlap between national and Union competences. Hence, in the texts and in practice, a significant focus has been put on coordination (both among national authorities, and between them and the Union authorities). Coordination issues arise in all areas where the Union was conferred a responsibility without the corresponding direct instruments; beyond fiscal policies, for which the problem is well-identified (see Chapter 3), this regards the structural policies that belong to the so-called *Lisbon Agenda** (see Chapter 6), especially Research and Development and labor markets, several regulatory domains such as banking supervision,

a large part of climate-change policies, and, increasingly, fields traditionally belonging to the realm of national decision, such as migration. However, while economic analysis provides arguments in favor of coordination and recommendations for implementing it, it also underlines its many difficulties (as already discussed).

Conclusion

There have been a few blessed decades, from the immediate aftermath of World War II to the early 1970s, when economic policy could be regarded as dedicated to the public good, conceptually simple, and reasonably easy to implement. When this golden age ended in the 1980s, policymakers and economists were cast out of the Garden of Eden. Since then, they have been living in a much more imperfect world.

The recognition of the limits of standard models and of the limitations implied by international interdependence should neither lead to underestimating the responsibility of policy nor to putting an excessive faith in the self-regulating virtue of markets. To determine what economic policy can achieve in this context, and on what conditions it can reach its goals, is the objective of the chapters that follow.

References

Alesina, A., and G. Tabellini (2007), "Bureaucrats or Politicians? Part I: A Single Policy Task," *American Economic Review*, 97, pp. 169–79.

Alesina A., and R. Wacziarg (1999), "Is Europe Going too Far?," *Carnegie-Rochester Conference Series one Public Policies*, Vol. 51, pp. 1–42, North Holland.

Alesina, A., I. Angeloni, and F. Etro (2005), "International Unions," *American Economic Review*, 95, pp. 602–15.

Anderson, J.E., and E. van Wincoop (2003), "Gravity with Gravitas: A Solution to the Border Puzzle," *American Economic Review*, 93, pp. 170–92.

Arrow, K. (1951), *Social Choice and Individual Values*, Wiley.

Arrow, K. (1968), "Optimal Capital Policy with Irreversible Investment," in J.N. Wolfe (ed.), *Value, Capital and Growth: Papers in Honor of Sir John Hicks*, Edinburgh University Press, pp. 1–19.

Axelrod, R. (1984), *The Evolution of Cooperation*, Basic Books.

Bachelier, L. (1900/2006), "Théorie de la spéculation" (Thesis), *Annales Scientifiques de l'École Normale Superieure* (1900) III-17, pp. 21–86. English translation (with an introduction) by M. Davis and A. Etheridge (2006) (foreword by P.A. Samuelson), *Louis Bachelier's Theory of Speculation: The Origins of Modern Finance*, Princeton University Press.

Baldwin, R.E. (2006), "The Euro's Trade Effects," ECB working paper no. 594, March.

Barro, R., and D. Gordon (1983), "A Positive Theory of Monetary Policy in a Natural Rate Model," *Journal of Political Economy*, 91, pp. 589–610.

Bernanke, B. (2002), "Deflation: Making Sure 'It' Doesn't Happen Here." Remarks before the National Economists Club, Washington, D.C., 21 November.

Bernanke, B. (2007), "Regulation and Financial Innovation", speech, Financial Markets Conference, Federal Reserve Bank of Atlanta, Sea Island, GA: 15 May (available at www.federalreserve.gov/boarddocs/Speeches/2007/20070515/default.htm).

Bernanke, B. (2008), *Testimony Before the Committee on Banking, Housing, and Urban Affairs*, US Senate, 14 February.

Bernoulli, D., (1738), *Specimen theoriae novae de mensura sortis*, Commentarii Academiae Scientiarum Imperialis Petropolitanae. Translated, Bernoulli, D. (1954), "Exposition of a new theory on the measurement of risk", *Econometrica*, 22, pp. 23–36.

Besson, D. (2002), " L'investissement des administrations publiques locales: Influence de la décentralisation et du cycle des élections municipales," *INSEE Première*, no. 867, October.

Black, D. (1948), "On the Rationale of Group Decision-making," *Journal of Political Economy*, 56, pp. 23–34.

Blinder, A. (1997), "Is Government Too Political ?," *Foreign Affairs*, November–December, p. 126.

Botman, D., Ph. Karam, D. Laxton, and D. Rose (2007), "DSGE Modeling at the Fund: Applications and Further Developments," *IMF Working Paper* no. 07/200, August.

Carlson, J., and N. Valev (2001), "Credibility of a New Monetary Regime: The Currency Board in Bulgaria," *Journal of Monetary Economics*, 47, pp. 581–94.

Coeuré, B., and J. Pisani-Ferry (2005), "Fiscal Policy in EMU: Towards a Sustainability and Growth Pact?", *Oxford Review of Economic Policy*, 21, pp. 598–617.

Collignon, S. (2002), "The European Republic; Reflections on the Political Economy of a European Constitution." The Federal Trust, London, England.

Deaton, A. (1992), *Understanding Consumption (Clarendon Lectures in Economics)*, Oxford University Press.

Dixit, A. (1996), "The Making of Economic Policy", Munich Lectures, The MIT Press.

Djankov, S., E. Glaeser, R. La Porta, F. Lopez-de-Silanes, and A. Shleifer (2003), "The New Comparative Economics," *Journal of Comparative Economics*, 31, pp. 595–619.

Epstein, L., and T. Wang (1995), "Intertemporal Asset Pricing Under Knightian Uncertainty," *Econometrica*, 62, pp. 283–322.

Erceg, C., L. Guerrieri, and C. Gust (2005), "SIGMA: A New Open Economy Model for Policy Analysis," Federal Reserve Board, *International Finance Discussion Papers* Number 835, July.

European Commission (2007), *Public Finances in EMU.*

Evans, G., and S. Honkapohja (2001), *Learning and Expectations in Macroeconomics*, Princeton University Press.

Feldstein, M. (1988), "Thinking about International Economic Coordination," Distinguished Lecture on Economics in Government, *The Journal of Economic Perspectives*, 2, pp. 3–13.

Fitoussi, J.-P. (2002), *La Règle et le Choix*, Le Seuil/La République des Idées.

Frankel, J., and K. Rocket (1988), "International Macroeconomic Policy Coordination when Policy Makers Do Not Agree on the True Model," *American Economic Review*, 78, pp. 318–40.

Ghosh, A., and P. Masson (1994), *Economic Cooperation in an Uncertain World*, Blackwell.

Gilardi, F. (2005), "The Institutional Foundations of Regulatory Capitalism: The Diffusion of Independent Regulatory Agencies in Western Europe," *The Annals of the American Academy of Political and Social Science*, 598, pp. 84–101.

Gollier, C. (2001), "Economie du principe de precaution," in F. Ewald, C. Gollier, and N. de Sadeleer (eds), *Le principe de précaution, Que Sais-Je No. 3596*, Presses Universitaires de France.

Gollier, C. (2002), "Time Horizon and the Discount Rate," *Journal of Economic Theory*, 107, pp. 463–73.

Gollier, C., B. Julien, and N. Treich (2000), "Scientific Progress and Irreversibility: An Economic Interpretation of the 'Precaution Principle'," *Journal of Public Economics*, 75, pp. 229–53.

Greif, A., P. Milgrom, and B. Weingast (1994), "Coordination, Commitment and Enforcement: The Case of the Merchant Gild," *Journal of Political Economy*, 102, pp. 745–76.

Grossman,G., and E. Helpman (1994), "Protection for Sale," *The American Economic Review*, 84, pp. 833–50.

Guesnerie, R. (2003), *Kyoto et l'économie de l'effet de serre*, La Documentation française.

Ha-Duong, M. (1998), "Quasi-Option Value and Climate Policy Choices," *Energy Economics*, 20, pp. 599–620.

Hayek, F. (1944), *The Road to Serfdom*, Routledge.

Henry, C. (1974), "Investment Decisions under Uncertainty: The Irreversibility Effect," *American Economic Review*, 64, pp. 1006–12.

Holmström, B., and J. Tirole (1989), "The Theory of the Firm," in R. Schmalensee and R. Willig (eds), *Handbook of Industrial Organization*, vol. 1, North–Holland, pp. 63–133.

Hotelling, H. (1929), "Stability in Competition", *The Economic Journal*, 39, pp. 41–57.

IMF and World Bank (2005), *The Standards and Codes Initiative—Is It Effective? And How Can It Be Improved?*, mimeo.

Jacquet, P., J. Pisani-Ferry, and L. Tubiana (2002), "Les institutions économiques de la mondialisation", in *Gouvernance mondiale*, report to the French Council of Economic Analysis (Conseil d'analyse économique), La Documentation française.

Kahneman, D. (2002), "Maps of bounded rationality: A perspective on intuitive judgment and choice," Nobel Prize Lecture, 8 December.

Keynes, J.M. (1936), *The General Theory of Employment, Interest and Money*. Reprinted in *The Collected Writings of John Maynard Keynes*, Vol. VII, 1972, Macmillan.

King, M. (2004), "The Institutions of Monetary Policy," *American Economic Review*, 94, pp. 1–13.

Knight, F. (1921), *Risk, Uncertainty and Profit*, Hart, Schaffner & Marx.

Krugman, P. (1998), "Japan's Trap," available on http://web.mit.edu/krugman/www/japtrap.html.

Kydland, F., and E. Prescott (1977), "Rules Rather than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, 85, pp. 473–91

Laffont, J.-J. (2000a), "Étapes vers un État moderne; une analyse économique", in *État et gestion publique*, Rapport no. 24 du Conseil d'analyse économique, La Documentation française.

Laffont, J.-J. (2000b), *Competition in Telecommunications*, MIT Press.

Laffont, J.-J. (2000c), *Incentives and Political Economy*, Clarendon Lectures, Oxford University Press.

Laffont, J.-J., and D. Martimort (2002), *The Theory of Incentives: The Principal-Agent Model*, Princeton University Press.
Laffont, J.-J., and J. Tirole (1986), "Using Cost Observations to Regulate Firms," *Journal of Political Economy*, 94, pp. 614–41.
Layard, R. (2005), *Happinness: Lessons from a New Science*, Penguin.
Lipsky, J. (2008), "Dealing with Financial Turmoil: Tail Risks, Policy Challenges, and the Role of the IMF,", speech given at the Peterson Institute, 12 March, Washington DC.
Lucas, R. (1976), "Econometric Policy Evaluation: A Critique," in K. Brunner and A. Meltzer (eds), *The Phillips Curve and Labor Markets*, Carnegie-Rochester Conference Series on Public Policy, vol. 1, pp. 19–46.
Mandelbrot, B., and R.L. Hudson (2004), *The (Mis)Behavior of Markets: A Fractal View of Risk, Ruin and Reward*, Basic Books.
Maskin, E., and J. Tirole (2004), "The Politician and the Judge: Accountability in Government ," *American Economic Review*, 94, pp. 1034–54.
Mayer, T, and G. Ottaviano (2007), *The Happy Few: The Internationalization of European Firms*, Bruegel Blueprint Series no. 3, Bruegel.
Mayer, T., and S. Zignago (2005), "Market Access in Global and Regional Trade," CEPII working paper 2005–02.
McCallum, J. (1995), "National Borders Matter: Canada–US Regional Trade Patterns," *American Economic Review*, 85, pp. 615–23.
McDonald, R, and D. Siegel (1986), "The Value of Waiting to Invest", *The Quarterly Journal of Economics*, 101, pp. 707–28.
McDougall, D. (1977), *The Role of Public Finance in European Integration*, mimeo, the European Commission.
Mishkin, F. (2008), "Monetary Policy Flexibility, Risk Management, and Financial Disruptions," Speech at the Federal Reserve Bank of New York, 11 January.
Musil, R. (1930), *The Man Without Qualities*, translated by E. Wilkins and E. Kaiser, Minerva, 1979.
Muth, J. (1961), "Rational Expectations and the Theory of the Price Movements," *Econometrica*, 29, pp. 315–35.
Nordhaus, W. (1975), "The Political Business Cycle," *The Review of Economic Studies*, 42, pp. 169–90.
Nordhaus, W. (2007), "A Review of the Stern Review on the Economics of Climate Change," *Journal of Economic Literature*, XLV, pp. 686–702.
Oates, W. (1972), *Fiscal Federalism*, Harcourt Brace Jovanovich.
Olson, M. (1969), "The Principle of Fiscal Equivalence: The Division of Responsibilities Among Different Levels of Government," *American Economic Review*, 59, pp. 479–87.
Oudiz, G. and J. Sachs (1984), "Macroeconomic Policy Coordination Among the Industrial Countries," *Brookings Papers on Economic Activity*, 1, pp. 1–64.
Persson, T. (1998), "Economic Policy and Special Interest Politics", *The Economic Journal*, 108, pp. 310–27.
Persson, T., and G. Tabellini (1990), *Macroeconomic Policy, Credibility and Politics*, Harwood Academic Publishers.
Persson, T., and G. Tabellini (1999), "Political Economics and Macroeconomic Policy," in J. Taylor and M. Woodford (eds) *Handbook of Macroeconomics*, North Holland.

Persson, T., and G. Tabellini (2001), "Political Institutions and Policy Outcomes: What are the Stylized Facts," *CEPR Discussion paper*, No. 2872.
Pisani-Ferry, J. (2004), "The Eurozone's Macroeconomic Framework: Does it matter? What should be done ?," in *Economic Reform in Europe: Priorities for the Next Five Years*, R. Liddle and M.J. Rodrigues (eds), Policy Network, November.
Posner, R. (2004), *Catastrophe: Risk and Response*, Oxford University Press.
Rogoff, K. (1984), "Can International Monetary Policy Coordination be Counterproductive?," *Journal of International Economics* 18, pp. 199–217.
Rosanvallon, P. (2000), *La démocratie inachevée, Histoire de la souveraineté du peuple en France*. Gallimard.
Rose, A.K. (2000), "One Money, One Market: Estimating the Effect of Common Currencies on Trade," *Economic Policy*, 30, pp. 9–45.
Salanié, B. (1997), *The Economics of Contracts: A Primer*, MIT Press.
Sapir, A., Ph. Aghion, G. Bertola, M. Hellwig, J. Pisani-Ferry., D. Rosati, J. Viñals, and H. Wallace (2004), *An Agenda for a Growing Europe*, Oxford University Press.
Shleifer, A., and R. Vishny (1993), "Corruption," *Quarterly Journal of Economics*, 108, pp. 599–617.
Sims, C. (1980), "Macroeconomics and Reality," *Econometrica*, 48, pp. 1–48.
Sinn, H.-W., and W. Ochel (2003), "Social Union, Convergence and Migration," *Journal of Common Market Studies*, 41, pp. 869–96.
Smith, A. (1776), *An Inquiry into the Nature and Causes of the Wealth of Nations*, Methuen & Co.
Stern, N. (2007), *The Economics of Climate Change*, Cambridge University Press and http://www.hm-treasury.gov.uk/independent_reviews/stern_review_economics_climate_change/stern_review_report.cfm.
Stern, N. (2008), "The Economics of Climate Change," Richard Ely Lecture, *American Economic Review*, 98, pp. 1–37.
Stigler, G. (1971), "The Economic Theory of Regulation," *Bell Journal of Economics and Management Science*, 2, pp. 3–21.
Stiglitz, J. (2000), "The Contributions of the Economics of Information to Twentieth Century Economics," *Quarterly Journal of Economics*, 115, pp. 1441–78.
Stiglitz, J. (2003), "Whither Reform? Towards a New Agenda for Latin America," *CEPAL Review*, 80, pp. 7–38.
Svensson, L. (2004), "Optimal Policy with Low-Probability Extreme Events," NBER Working Paper, no. 10196.
Tabellini, G. and Ch. Wyplosz (2004), *Réformes structurelles et coordination en Europe*, La Documentation française.
Tiebout, C. (1956), "A Pure Theory of Local Expenditures," *The Journal of Political Economy*, 64, pp. 416–24.
Tobin, J. (1990), "On the Theory of Macroeconomic Policy," *De Economist*, 138, pp. 1–14.
Tucker, A. (1950, 1980), "A Two-Person Dilemma," conference given at Stanford University, later published as "On Jargon: The Prisoner' s Dilemma," *UMAP Journal*, 1, p. 101.
Von Neumann, J., and O. Morgenstern (1944), *The Theory of Games and Economic Behavior*, Princeton University Press.
Wallison, P. J. (2007), "Fad or Reform. Can Principles-Based Regulation Work in the United States?," Financial Services Outlook, AEI Online, Washington DC: American Enterprise Institute (8 June) (available at www.aei.org/include/pub_print.asp?pubID=26325).

Weber, M. (1919), "Politics as a Vocation," in *Max Weber: Selections in Translation.* Cambridge University Press, 1978.
Weingast, B. (1995), "The Economic Role of Political Institutions: Market-preserving Federalism and Economic Development." *Journal of Law, Economics, and Organization*, 11, pp. 1–31.
Weitzman, M. (2007), "A Review of the Stern Review on the Economics of Climate Change," *Journal of Economic Literature*, XLV, pp. 686–702.
Yi, K.M. (2010), "Can Multi-Stage Production Explain the Home Bias in Trade," *American Economic Review*, 100, pp. 364–93.

3

Fiscal Policy

3.1 Issues
  3.1.1 What is it all about?
  3.1.2 Lessons from history
3.2 Theories
  3.2.1 Demand-side effects: Keynes and his critics
  3.2.2 Public debt sustainability
  3.2.3 Supply-side effects and reconciliation attempts
3.3 Policies
  3.3.1 Rules and principles for fiscal policy
  3.3.2 Fiscal policy in the European monetary union
  3.3.3 Discretionary fiscal policy in times of crisis
References

The public budget, which includes the budgets of central and local governments and for social insurance, simultaneously fulfils the three functions of allocation, redistribution, and stabilization analyzed in chapter 1. Even though these functions are intricately intertwined (Buiter, 1990), the notion of fiscal policy usually refers to the stabilization function, and can be defined as the set of decisions or rules regarding taxes and public expenditures for purposes of dampening the fluctuations of the economic cycle in order to keep unemployment close to its equilibrium value and avoid the build-up of deflationary or inflationary pressures (Samuelson, 1948).

Under this definition, fiscal policy emerges as a twentieth-century invention that owes considerably to the thinking of John Maynard Keynes—even though the history of public spending and of its financing is obviously much older. But it owes even more to the general rise of the share of public expenditures in GDP as a consequence of the generalization of government-financed social insurance, welfare, and education. In the US, federal expenditures as a proportion of GDP rose from 2–3% before World War I to 5% in the 1920s, 10% in the 1930s, 15% in the aftermath of World War II, and then stabilized at around 20% in the 1960s. Within the course of half a century, the federal government has thus been transformed from an irrelevant macroeconomic player into a major contributor to aggregate demand (figure 3.1). Similar evolutions have been observed in other countries.

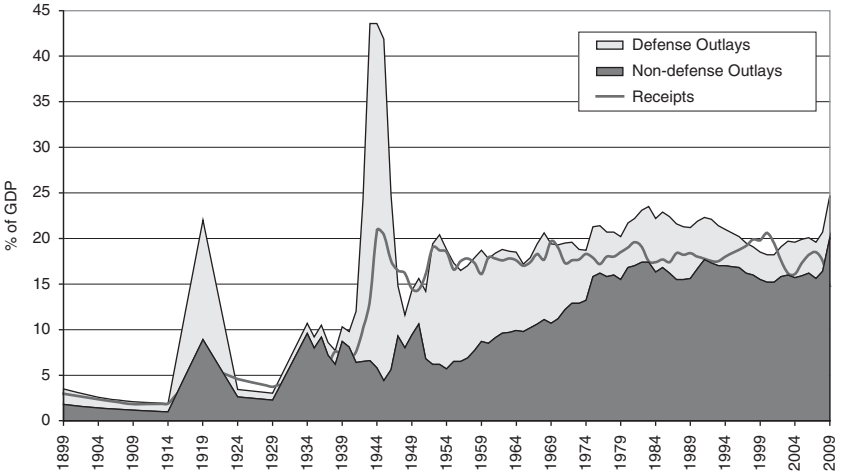


Figure 3.1 US federal expenditures and outlays as percentage of GDP, 1899–2009. Source: Bureau of Economic Analysis, historical statistics up to 1970; Economic Report of the President (2010) since 1970.

Toward the end of the twentieth century, theoretical and empirical doubts surfaced about the effectiveness of fiscal policy as a stabilization tool. Experiences with failed fiscal expansions or painless consolidations in several countries as well as policy–philosophy reversals have prompted a reconsideration of old issues and exploration of new ones. Are fiscal expansions effective, particularly when public debt reaches a high level? Conversely, does fiscal contraction always have a recessionary effect on demand? Is it possible and desirable to conceive, and abide by, fiscal policy principles and rules? How should relations between various levels of government or between members of a monetary union be organized? Who eventually pays the public debt? After 2008, further questions came to the fore in the wake of the financial and economic crisis, as fiscal policy was rehabilitated as a key tenet of the policy response: How big should fiscal stimulus packages be? Should they rely on tax cuts or spending increases? For how long should they be maintained? What is the desirable exit strategy? How to gear a timely and sizeable fiscal expansion while maintaining public debt sustainability?

3.1 Issues

3.1.1 What is it all about?

a) What is a budget?

A public *budget** is a document that specifies the origin and volume of both income ("receipts") and its intended spending over a certain time horizon

(usually a year). Receipts consist of income from direct and indirect taxation, social insurance contributions, revenues from—and possibly disposal of—public assets or sale of public services. Spending is made on activities such as defense, police, justice, education, research, support to the economy, social policy, health, foreign policy, development assistance, etc. Budgets are drafted at different levels of government, from municipalities to central governments, but the stabilization function is usually mainly shouldered by the central government.

The preparation of the draft budget, its discussion and its adoption by the relevant legislative body are important stages of economic-policy decision-making. Organization, procedures, and the time frame vary substantially from country to country.[1] A typical sequence includes: The definition of the overall macroeconomic framework leading to forecasts about government receipts; the setting of expenditure ceilings by sectors; the preparation, by the government departments, of their own draft budgets; cabinet discussions and the consolidation of the whole budget; discussions and the vote in parliament. The overall process requires at least six months.

As an example, the federal budget process in the US begins with the submission of the President's budget proposal to the US Congress, on the first Monday of February each year. By early April, the House and Senate budget committees (which may or may not draw on the President's proposal) submit their draft resolutions to the floor for adoption. Then a joint Senate and House conference report is prepared to reconcile the two versions and a concurrent budget resolution is adopted that identifies categories of spending (for instance, national defense, education, agriculture, etc.). In each spending category, the distinction is made between mandatory and discretionary spending. The first refers to spending that is not subject to current Congressional approval (for example, it may result from entitlements and other effects from laws enacted in the past—*services votés* to use the telling French expression). Discretionary spending, however, requires an *appropriation bill** through which Congress authorizes the President to commit and spend resources.

In many countries, fiscal policy is conducted within a medium-term framework which specifies the yearly evolution of expenditures, or establishes debt and deficit ceilings. Within the European Union, for example, all member states of the euro area have to submit *stability programs** that briefly describe the planned three-year evolution of major budget components and their sensitivity to alternative growth scenarios.[2] Independently of any international obligation, some countries have adopted internal, more-or-less-binding fiscal rules, such as the balanced budget rules adopted by several US states, the

1. In Japan and the UK, for example, the financial year runs from 1 April to 31 March; in the US, from 1 October to 30 September; in France, the state budget is drawn up on the basis of the calendar year.

2. EU member states outside the euro area are subject to a comparable procedure.

*golden rule of public finance** adopted by Germany in the 1970s and by the UK in the 1990s stipulating that, in principle, only investment expenditures can be financed through debt, or the obligation, enshrined in the German constitution in 2009, to limit the federal deficit to 0.35% of GDP over the cycle starting in 2016 (see section 3.3.1).

Because a large proportion of the budget is devoted to civil servant compensation and pensions, and core government missions such as security and justice, and because some expenditure categories (infrastructure, defense) are subject to multi-year programming, the room for maneuver for fiscal policymakers is generally limited in the short run, which makes it difficult, for instance, to rapidly reduce public indebtedness, unless by selling government assets.

Additionally, given the length of the decision process, fiscal policy is difficult to use for counter-cyclical purposes, especially when decisions have to be taken outside the normal yearly budgetary process. In fact, both monetary and fiscal policies affect economic activity after a lag, but for different reasons. The impact of monetary policy is delayed due to fixed-rate indebtedness of households and firms, imperfect reaction of long-run interest rates, or lagging reaction of the banking sector. Conversely, fiscal policy has immediate impact on demand through public consumption and investment, or through households' disposable income, but the fiscal decision process is much longer than the monetary one because it requires several instances of negotiation within the government and with parliament. While some models treat fiscal and monetary policies in similar ways, these policies neither have the same flexibility nor the same reactivity.

The *fiscal (or budgetary) balance** is the difference between income and expenditure. It may be calculated for a specific segment of government (central government, state and/or local authorities, the social insurance entity) or for the *general government**, which consolidates all government accounts. Fiscal balance can also be calculated by excluding some categories of expenditures. Importantly, the *primary balance** excludes interest payments on public debt (cf. box 3.1); the UK government also publishes a current fiscal balance that excludes public investment spending.

There is a *fiscal (or budget) surplus** when the budget balance is positive, and a *fiscal* (*or budget*) *deficit** when the balance is negative. Surpluses can be used to pay down the public debt, or are invested. Several governments have established *sovereign wealth funds** which acquire foreign assets and are funded either by budget surpluses or by the transfer of foreign exchange reserves from the central bank (see chapter 5). Examples include Singapore (*Government of Singapore Investment Corporation*, or GIC), Abu Dhabi (Abu Dhabi Investment Authority or ADIA), and Norway (where oil income is invested in the *Government Pension Fund of Norway*).

Although there has been a movement toward decentralization in a number of industrial as well as developing countries, the degree of fiscal decentralization, or, conversely, of fiscal federalism, still varies substantially

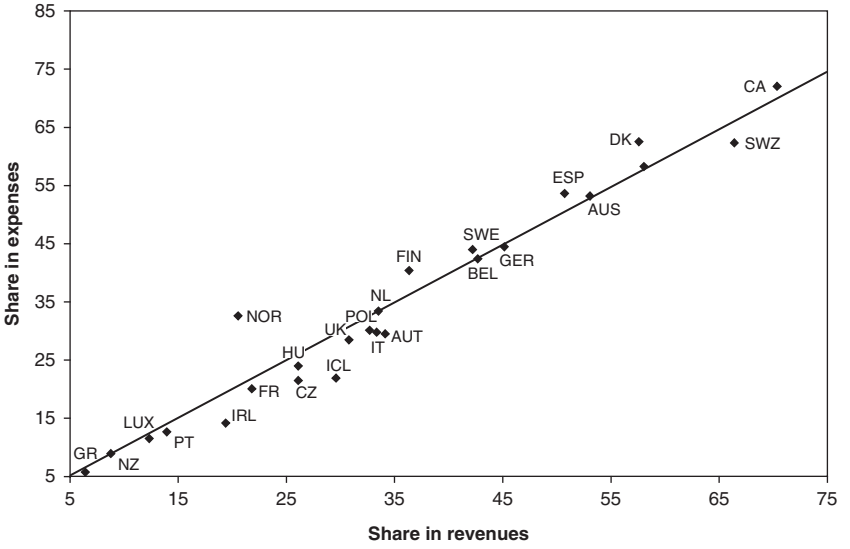


Figure 3.2 Ratio of local to general government expenses and revenues (2007 or closest, in %).
Source: International Monetary Fund, *Government Finance Statistics*, December 2009.
Note: Regional/state expenditures are included when relevant.

across countries (see figure 3.2). The ratio of state and local to general government expenditures ranges from 5% in Greece to 70% in Canada.

b) Deficit finance

Leaving out the option of selling assets, deficits need to be financed, either by borrowing from the national central bank, which amounts to creating money, or by borrowing (or, in the case of poor developing countries, by receiving grants) from other public and private agents, including international organizations or foreign governments. Development assistance will not be discussed in this chapter but can play a substantial role in less-developed countries. In 2006, grants and aid related to debt forgiveness contributed to an average 1.3% of GDP to finance the budget of Western African Economic and Monetary Union countries, and were as large as 5.7% of GDP in Guinea-Bissau.[3]

The *monetization of the deficit** consists of an overdraft or a loan granted by the central bank to the government that increases the money supply. This practice originates in the capacity of kings to finance their expenses by printing money and cheating on the actual weight of gold coins (seigniorage, see chapter 4) and it used to be common in the past, in particular to finance wars.

3. Source: Western African Economic and Monetary Union (2007).

It amounted to letting the requirements of the public Treasury, rather than the economy's transaction or hoarding needs, determine the pace of money creation; and, when sustained over time, it is a powerful source of inflation (see the Sargent and Wallace model introduced in chapter 4, box 4.11). For example, hyperinflation episodes in Latin America were systematically connected to devious public finance. With less than full indexation, money creation and inflation provided governments with a way to escape the debt burden, ultimately shifting the burden from the taxpayer to the money-holder and the consumer. Jacques Necker who, in his capacity as King Louis XVI's Director General of Finance had gotten France into massive debt in order to finance the American War of Independence, had recognized some aspects of this mechanism:

> One needs to keep in mind an important truth, namely that, without any effort, and by virtues of nature, the burden of the public debt diminishes every day. A given nominal amount will not be worth twenty years from now, if one is allowed this comparison, what it is worth today, because its relation to the price of all goods will necessarily change with the progressive increase in gold and silver: time therefore contributes to the amortization of public debt.
>
> J. Necker (1784), p. 113, authors' translation

This link between deficit finance and inflation has led to explicit or implicit restrictions on how governments can borrow from their central bank. It is, for example, the reason why euro area Treasuries are forbidden to seek funding from the European Central Bank or any of the national central banks.[4] Such restrictions are now widespread. Hence, public deficits need to be financed in other ways, at least in normal times. The issue of debt (deficit) monetization typically resurfaces in the midst of sharp financial crises that spill over to the real economy. Faced in 2003 with a large-scale economic crisis and with the risk of a deflationary spiral, the Bank of Japan undertook to monetize part of the public debt in order to fuel new expectations of inflation. However, it did so by buying Government securities on the market rather than by direct lending to the Japanese Treasury. In the wake of the 2007 subprime crisis and 2008 financial collapse, the Federal Reserve and other central banks, after having pushed interest rates down through more conventional means, undertook to buy long-term Treasury securities, thus signaling their willingness to monetize the debt in order to further ease the liquidity strain and spur aggregate demand.[5]

In advanced economies, *public borrowing** consists in selling to investors debt securities giving them the right, for a given period of time, to payments in capital and interest specified by the associated debt contract (box 3.1). In many emerging economies, governments also borrow from banks and

4. Each Treasury's cash account with its national central bank has therefore to be in surplus at the close of every business day.

5. This move is part of the strategy of "quantitative easing" (see chapter 4).

from international institutions such as multilateral development banks. Accumulated borrowing constitutes *public debt**. Public debt represents the financial liabilities of the public sector vis-à-vis private actors. It should not be confused with external debt, which represents the liabilities of all domestic actors vis-à-vis the rest of the world. Of course, both concepts have a common component: Government bonds purchased by nonresident investors, which represent more than 50% of US Treasuries and more than 60% of French and German government bonds, are part of both public and external debt.

Box 3.1 The Market for Public Debt

In advanced economies, governments fund their financing needs by issuing securities called Treasury (or government) bills and bonds. The amount issued within a given year has to cover the deficit of the year and the reimbursement of debt coming to maturity. This task is performed by ministries of finance or by separate agencies called debt management offices. Debt securities can be either short-run (*Treasury bills**, for example with a three-month or a one-year maturity) or long-run (*government bonds**, up to a 50-year maturity). The interest rate paid on debt is usually fixed but in some instances it can be variable. In particular, some countries issue inflation-protected bonds which pay a fixed real interest rate. The average maturity of major countries' public borrowing typically lies between 5 and 10 years but it is somewhat shorter in the US and longer in the UK. OECD countries generally borrow in their domestic currency, while emerging countries with less-developed financial markets often borrow in US dollars, and to a lesser extent in euros. Government bonds are traded by investment banks and eventually purchased by institutional investors such as asset managers and pension funds (which manage household savings), insurance companies, central banks, and sovereign wealth funds.

Unless they are in financial distress, governments are usually considered more solvent than any private agent, and the interest rate they pay on their debt is thus considered as the *risk-free interest rate** and serves as the basis on which all financial securities are valued. However, not all governments are equal and their borrowing costs depend on their *credit quality**, i.e., the likelihood assigned by investors that they could become insolvent in the future. Investors often rely on the opinions expressed by *rating agencies**, independent institutions which assess the creditworthiness of borrowers. They will tend to apply a *risk premium** that depends on the rating of each government and will require different interest rates according to such ratings. This is apparent in figure B3.1.1 which describes the public debt financing cost of France, Greece, Ireland, Italy, and Spain relative to that of Germany from May 2007 to April 2010. Interest-rate spreads on ten-year government bonds between countries remained rather narrow

(amounting at most to 30 basis points per annum, that is 0.3%, between the German and Greek interest rates) until the second quarter of 2008. Spreads for Greece, Ireland, Portugal and, to a lesser degree, Italy widened considerably during the fall of 2008 and the beginning of 2009, a sign of investors' mistrust in the ability of heavily indebted governments to keep public finance on a sustainable path during the economic crisis. In the spring of 2010, following serious doubts about the ability of Greece to stabilize its debt-to-GDP ratio and about the ability and willingness of its European partners to provide adequate support, spreads on Greece jumped to over 300 basis points.

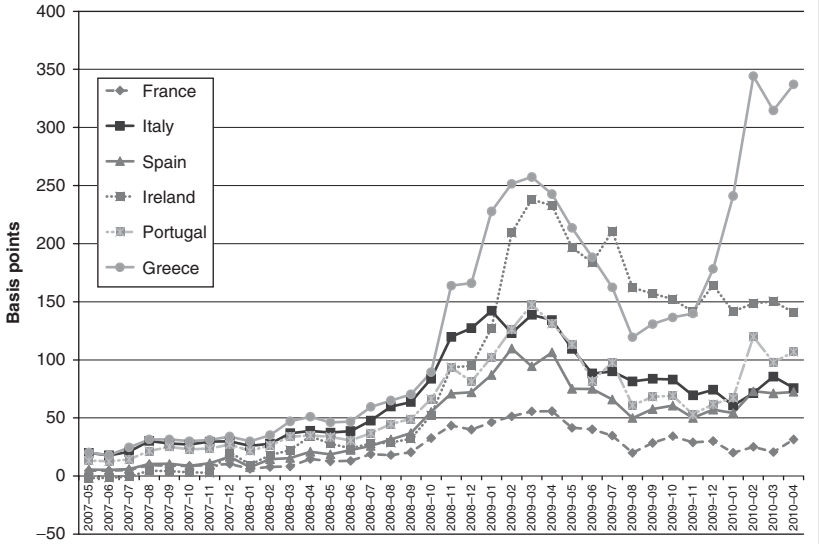


Figure B3.1.1 Credit discrimination among euro area governments: spreads on 10-year government bonds (in basis points).
Source: Thomson Datastream.

Central banks typically hold Treasury bonds as one of the counterparts of money; they buy (or accept as collateral in repurchase agreements, see chapter 4) these securities from banks in exchange for providing liquidity. This mechanism differs from outright monetization of the deficit, since the central bank is not mandated by the government to buy or sell these securities and the amounts derive from monetary policy, not fiscal policy considerations.

Debt-financed public spending may still invite indirect or *ex post* monetization. For example, if the central bank aims at stabilizing the interest rate, a debt-financed fiscal deficit will induce money creation. A government whose

debt, held in the form of fixed-interest bonds, is perceived to be too high is often tempted—when the central bank is not fully independent—to engage in inflationary policies that will in effect devalue the debt and reduce the real value of the debt service (capital and interest). Moreover, international finance has displaced the monetization of deficits. For instance, in the 2000s, securities issued by large countries' governments to finance their deficits were acquired notably by central banks in Asia to prevent their domestic currencies from appreciating (which increased their foreign exchange reserves: This mechanism is detailed in chapter 5), leading to de facto monetization by current account surplus countries.[6]

c) Measuring the fiscal imbalance

The elaboration of relevant statistics about fiscal balances (surpluses or deficits) requires choices regarding both the institutional perimeter and the type of income and expenditures to be considered. Such choices will be dependent on the kind of information that is looked for.

Central or general government? The most widespread concept of fiscal balance focuses on the general government balance that consolidates central government, local governments, social insurance, and, when appropriate, federal states. This is a coherent perimeter as it includes all agents whose income mainly comes from tax payments and mandatory contributions, while allowing for different degrees of decentralization as illustrated in figure 3.2. Most international comparisons rely on this concept.

However, focusing on the general government may not always be relevant in terms of the budgetary process or from a political economy perspective, since the responsibility and the decision-process differ depending on the level of government (central, local, state, social insurance). Admittedly, only aggregate data determine the tax burden supported by the taxpayer. However, the use of aggregate figures for the public sector as a whole can obfuscate differences between contrasted situations at the sub-sector level.

Generally, most of the fiscal imbalance that occurs in general government arises from the central government (see figure 3.3). However, the central government may sometimes substitute for other levels of government. In Japan, for instance, the central government is highly indebted, but public pension funds have accumulated assets to prepare for the payment of future retirements. Would the state be able to draw from the accumulated surpluses in social accounts in the event of a difficulty? When asset or debt transfers between sub-sectors are difficult to implement, only a disaggregated approach can decipher the genuine dimension of potential problems. In several countries, notably in the US, local authorities have been allowed

6. Moreover, when reserve accumulation was not sterilized (see chapter 4), this also led to an expansion of the domestic money supply in Asian countries.

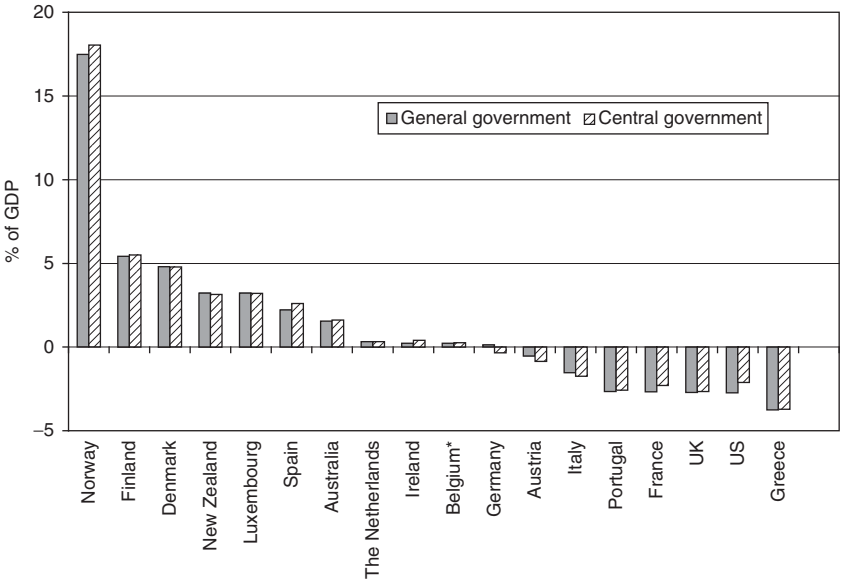


Figure 3.3 General and central government balances in 2007 (% of GDP). *Belgium: 2006.
Source: IMF, Government Finance Statistics, December 2009.

to go bankrupt without financial solidarity between various segments of government being called into play: It is indeed one of the characteristic features of the US model of fiscal federalism.

Total (financial) or primary deficit? Total fiscal balance, also called *net lending** or *financial balance**, is the difference between the expenditure of the public sector and its income. It represents the borrowing need of the government.[7] The financial balance includes the interest paid on public debt. For example, the Belgian and Italian governments had to pay more than 10% of GDP as interest charges on the public debt in the early 1990s. Interest charges depend on the debt level and on long-term interest rates, two variables that, in the short run, are not in governments' hands. A better indicator of

7. Here, there is an accounting subtlety: published fiscal balances are usually measured on an accrual basis, meaning that they register all operations that are *decided* in the year, irrespective of the date they are cashed into or out of the government's bank account. The borrowing need of the government is a cash concept and it is slightly different. In the UK, the former concept is called the *public sector net borrowing* and the latter is called the *public sector net cash requirement*. In all cases, however, it should be remembered that bond and bill issuance is usually larger than net lending because of the need to redeem debt maturing during the year, although the borrowing need can also be diminished by privatization proceeds.

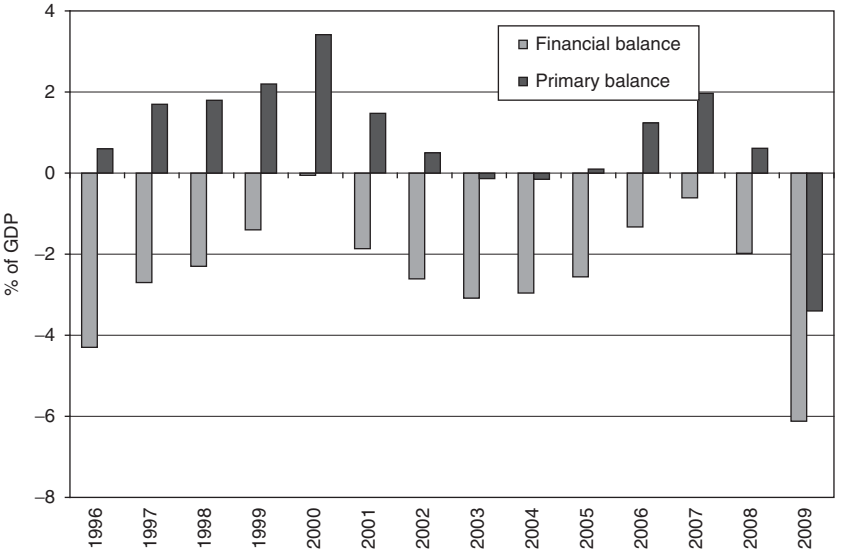


Figure 3.4 General government financial and primary balances in the euro area, 1996–2009.
Source: *OECD Economic Outlook* no. 86, November 2009.

the deliberate fiscal action of government and of parliament is the primary balance, defined as financial balance excluding interest payments:

$$\text{Financial balance (net lending)} = \text{primary balance} - \text{interest payments}$$

Figure 3.4 illustrates the difference between financial and primary balances in the euro area between 1996 and 2009. Interest payments on general government debt have accounted for close to 5% of GDP on average in the 1990s.

In emerging countries, public debt often has a short maturity, and interest rates, which reflect markets' expectations regarding the probability of failure of the borrowing government, are very unstable. For this reason, the primary balance is generally viewed as a more workable measure of aggregate fiscal policy. It is also key to understanding the dynamics of debt, as we shall see shortly.

Actual (financial) or cyclically adjusted (structural) deficit? A general pattern of fiscal balances is that they tend to rise when economic activity booms and to decline when it is slowing down. This is because most tax bases move in line with economic activity (for instance, VAT revenues depend on final consumption) whereas some components of public spending (e.g., unemployment benefits) slow down in economic booms. This spontaneous variation of fiscal balances—known as the *automatic stabilizers**—has a stabilizing effect on households' aggregate income since taxes paid, net of social transfers, increase

during economic expansions, while the reverse occurs during downturns, without any policy change.

In order to capture changes in fiscal policy, it is therefore useful to calculate a *cyclically adjusted balance** (also called *structural balance**) that measures what the financial balance would be, should output be at its potential level (cf. box 3.2). The change in the cyclically adjusted balance from one period to next is generally regarded as providing a measure of the *discretionary** component of fiscal policy because, in contrast to changes resulting from the automatic stabilizers, it results from a government decision. The evolution of the financial balance thus decomposes into a cyclical component, independent of the government's will, and a discretionary component, equal to the variation of the structural balance. The discretionary component provides a measure of the *fiscal stance**, i.e., of the orientation of fiscal policy.

$$\begin{aligned}\text{Financial balance (net lending)} &= \text{cyclical balance} + \text{cyclically adjusted balance} \\ &= \text{cyclical balance} + \text{structural balance}\end{aligned}$$

This measure of fiscal stance is the main indicator used by economists to shed light on policy debates, notably in the EU (see section 3.3). However it raises a host of technical debates related to the difficulty of measuring the output gap and the elasticity of government expenditures and receipts to the level of economic activity. For example, the change from one year to the next in the cyclically adjusted balance is meant to represent discretionary policy actions, but it often does not match estimates based on actual decisions regarding tax and spending—the difference sometimes being wide. Estimates by national governments and international organizations also differ, sometimes widely, and are subject to significant revisions over time. Therefore, the concept of structural balance is an important one for policy discussions, but estimates are far from being perfectly reliable guides for policy decisions.

Box 3.2 Calculating the Structural (Cyclically Adjusted) Public Balance

The structural (or cyclically adjusted) public balance is the public balance that would obtain had GDP been at its potential level. To calculate it, the first step is to assess the position of the economy in the business cycle, as measured by the output gap, i.e., the divergence of production y from its potential level $\bar{y}$ (y and $\bar{y}$ being in logarithm). Then, it is necessary to estimate, from past observations, the average sensitivity of the financial balance s, measured as a percentage of the GDP, to a variation of the output gap:

$$\varepsilon = \frac{ds}{d(y - \bar{y})} > 0 \qquad \text{(B3.2.1)}$$

The final step is to subtract the cyclical component $\varepsilon(y - \bar{y})$ from the financial balance s to get the cyclically adjusted, or structural, balance s^*:

$$s^* = s - \varepsilon(y - \bar{y}) \qquad \text{(B3.2.2)}$$

The measure of s^* naturally depends on the method used to calculate potential output (cf. chapter 1) and on the estimation of ε which can alternatively be performed on aggregate or disaggregate fiscal data. ε is thought to be close to 0.5 in the four major euro area countries (Germany, France, Italy, Spain), and close to 0.7 in Finland and 0.8 in The Netherlands (cf. Buti and Sapir, 1998, p. 132). When $\varepsilon = 0.5$, a 1% (of potential GDP) decline of the output gap mechanically raises financial balance by about 0.5% of GDP.

Structural balances have shortcomings of their own: Potential output is a notoriously fragile notion; economically relevant tax elasticities should be computed at a much disaggregated level, which is usually not the case, and they are unstable over time (figure B3.2.1).

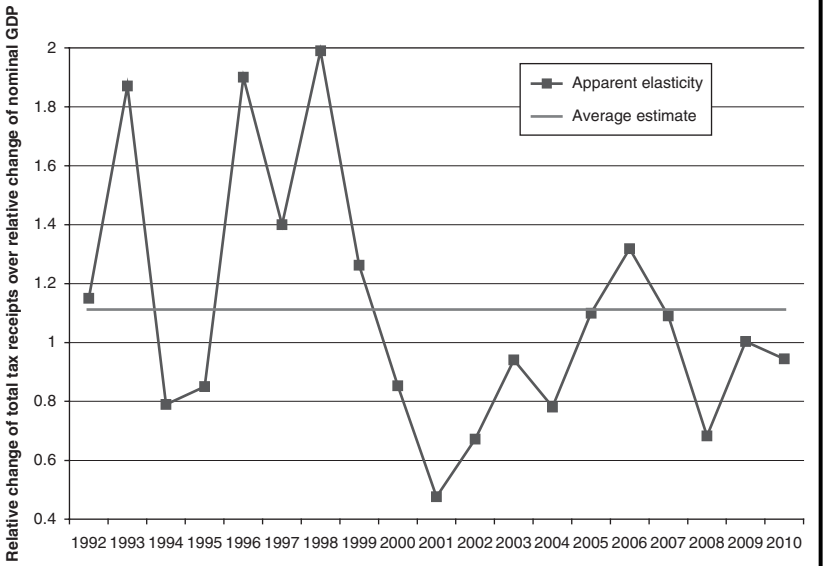


Figure B3.2.1 Euro-area implicit tax-elasticity with respect to GDP.
Source: Authors' calculations based on AMECO (April 2009, for 1999–2010, estimates for 2008, projections for 2009 and 2010) and European Commission (2006, p. 98, for 1992–98).

It can be useful to combine the two decompositions of the deficit (financial/primary, financial/structural) to calculate a *structural primary balance**. Since the interest on the debt is not very cyclical (because

governments mostly borrow at a fixed interest rate), one can write:

$$\begin{aligned}\text{Financial balance (net lending)} = {} & \text{cyclical primary balance} \\ & + \text{structural primary balance} \\ & - \text{interest payments on the debt}\end{aligned}$$

Like the financial balance, however, primary and structural balances include a number of nonrecurrent, large one-off fiscal operations such as privatization proceeds. These one-off operations undermine the accuracy of structural balances as indicators of the fiscal stance (see for example Joumard et al., 2008). For that reason, the OECD has introduced in 2008 a new indicator, the *underlying fiscal balance**, which measures cyclically adjusted fiscal deficits adjusted for one-off operations. In the same spirit, the OECD also publishes underlying primary fiscal balances. The above relation thus becomes:

$$\begin{aligned}\text{Fiscal balance (net lending)} = {} & \text{cyclical primary balance} \\ & + \text{one-off operations} \\ & + \text{underlying primary balance} \\ & - \text{interest payments on the debt}\end{aligned}$$

Figure 3.5 illustrates the usefulness of these decompositions in the case of the euro area. The graph indicates an improvement of both the financial and the primary balances from 1994 to 2000. This improvement was led by a rise in the underlying primary balance from 1994 to 1997 (in line with the need to abide by the Maastricht convergence criteria), whereas from 1998 to 2000 the improvement in the financial balance was essentially cyclical. Starting in 2007, there has been a marked cyclical deterioration of fiscal balances, as a result of the collapse of revenues and, to a lesser extent, of discretionary fiscal stimulation programs.

d) Public debt

Like private companies (but unlike households), the public sector need not repay its debts entirely because it is not expected to die. If debt grows too rapidly, however, investors who buy debt securities may become concerned about the future capacity of the government to raise new financing; hence some doubts may arise about the solvency of the public sector. As seen in box 3.1, such doubts may push up the interest rate at which the government borrows. However, the same rate of debt accumulation will not have the same meaning in a low-growing country as in a fast-growing country, because the capacity of the government to raise taxes broadly depends on nominal GDP. Therefore, public debt is generally measured as a ratio to GDP. As detailed in box 3.3, the same primary deficit leads to faster debt accumulation the higher the real interest rate and the lower the GDP growth rate. When the growth rate is higher than the interest rate, a country can stabilize its debt

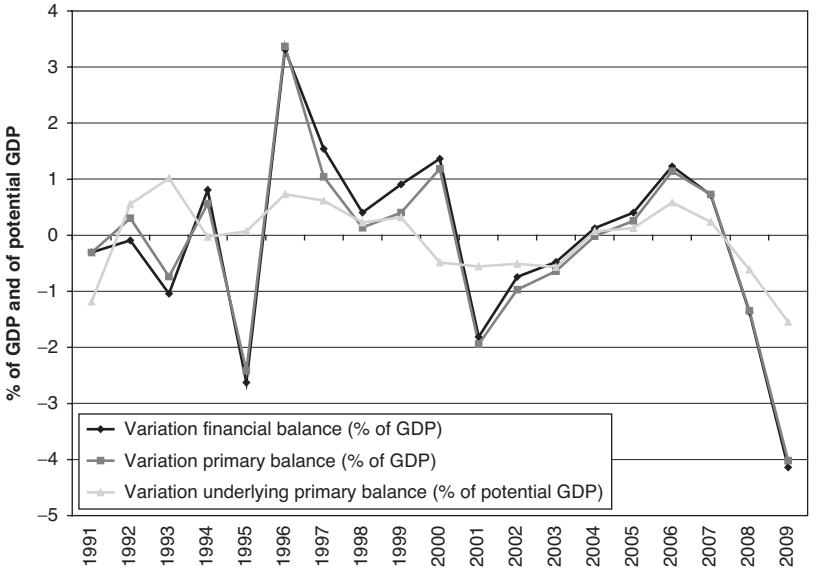


Figure 3.5 Fiscal balances of the euro area.
Source: *OECD Economic Outlook* no. 86, November 2009.

ratio even while maintaining a permanent primary deficit. Conversely, when the interest rate is higher than the growth rate, there must be a primary surplus to stabilize the ratio of debt to GDP; and the larger the (positive) difference between the interest rate and the growth rate, the larger the necessary primary surplus.[8]

Figure 3.6 provides an illustration of this arithmetic: In the 1990s, the US and France both experienced large fiscal deficits; but the debt-to-GDP ratio increased continuously in France while it stabilized in the US. The reason why deficits of similar relative magnitude in France and the US did not result in the same debt dynamics is that the US growth rate was higher than the French rate.

Like for private companies, the same debt dynamics may not have the same meaning, depending on what the borrowing resources are used for. For instance, financing new infrastructures may not worsen the long-term fiscal position of the government, for two reasons. First, additional infrastructure

8. These various elements are interdependent. For example, if a country's policy results in primary surpluses, declining risk premiums on government bonds may result in a fall in the interest rate. Italy benefited from such virtual dynamics in the 1990s when its government engineered a fiscal entrenchment to meet the Maastricht criteria for joining the European Monetary Union. These efforts resulted in a dramatic fall in Italian interest rates, which further accelerated deficit reduction. Symmetrically, the large primary deficit of Greece in 2009 (8% of GDP) triggered a rise in the risk premium that worsened the outlook of its public finances.

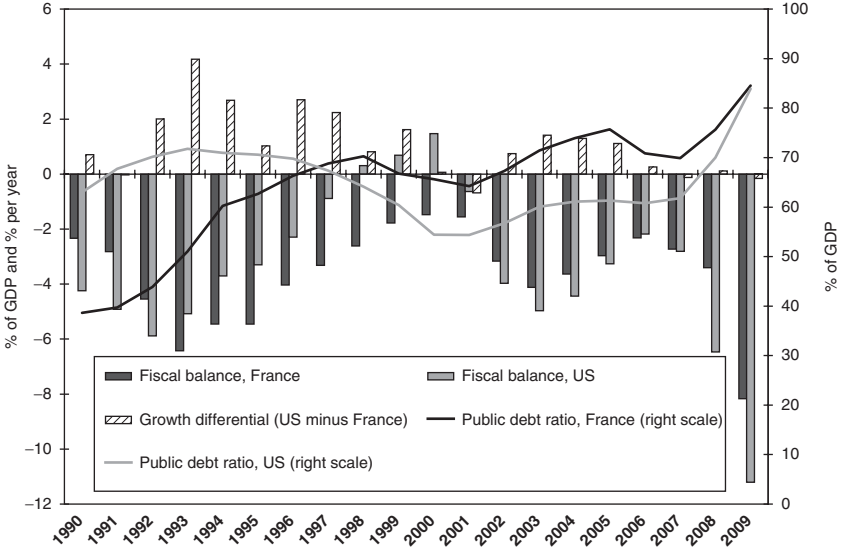


Figure 3.6 Debt dynamics in France and the US.
Source: *OECD Economic Outlook* no. 86, November 2009.

Box 3.3 Public Debt Dynamics

Let us denote as D the primary public deficit and B the public debt at year-end, both in euros, and i the nominal interest rate. We neglect cash revenues or disbursements (such as asset sales and purchases) that may impact public debt for a given public deficit. We also suppose that debt is measured at face value and not at current market value, thus ignoring valuation effects. Such assumptions are not innocuous: In emerging countries, part of the public debt is US dollar-denominated, and exchange-rate movements impact on debt dynamics (see chapter 5).

Indexing by -1 the values of the preceding period, the debt dynamics can be written as:

$$B = (1 + i)B_{-1} + D \qquad \text{(B3.3.1)}$$

Denoting as d and b, respectively, the primary deficit and the debt ratio as a percentage of nominal GDP, n the nominal growth rate (growth in volume + inflation), g the real growth rate, π the rate of inflation and r the real interest rate, we have:

$$n = g + \pi \qquad \text{(B3.3.2)}$$

$$\text{and} \quad i = r + \pi \qquad \text{(B3.3.3)}$$

Debt dynamics can thus be expressed as:

$$b = \frac{(1+i)}{(1+n)} b_{-1} + d \cong (1 + i - n) b_{-1} + d$$

$$\cong (1 + r - g) b_{-1} + d \qquad \text{(B3.3.4)}$$

or, equivalently:

$$b - b_{-1} = b_{-1}(i - n) + d = i b_{-1} + d - n b_{-1} \qquad \text{(B3.3.5)}$$

The variation of the debt ratio breaks up into three components: Interest payments on past debt, the primary deficit, and a relative diminution of the debt ratio through nominal growth. Two countries with similar primary deficits d will experience different dynamics depending on their real interest rate r compared to their real growth rate g, or equivalently, on their nominal interest rate i compared to their nominal growth rate n.

may raise GDP growth and hence curb the future debt ratio. Second, public infrastructures are assets that may be sold if necessary at a later point in time. This second reason suggests another way to assess public debt: By comparing it to public assets. Table 3.1 provides calculations of the *net public debt** (i.e., the difference between the *gross public debt** measured at market value and the value of public assets) for a few countries. Unsurprisingly, the net debt ratio is generally much lower than the gross one. It is sometimes even negative, meaning that public assets exceed public debts.

The use of net public debt is, however, debatable, since a number of public assets cannot be sold. The Japanese government, for example, can sell its shares in the Japan Post, but will have more difficulty selling the golden shrine in Kyoto. Net debt ratios are therefore partial images of the government's financial position and they tend to give an unduly favorable image of its financial situation.

Several governments are now producing comprehensive financial statements which provide the general public with a more faithful image of their financial situation. These comprise a description and valuation of government assets and liabilities and of some off-balance-sheet claims and liabilities. Table 3.2 summarizes the results for the US federal government.[9] But there are several reasons why governments cannot fully emulate private sector financial reporting. First, governments' primary function is not to sell goods and services and they cannot terminate their operations overnight. It is therefore unclear whether their balance sheet should be evaluated on the basis of market or historical prices. Second, many government assets are intangible

9. For other countries, see the references at the end of this chapter.

Table 3.1
Gross and net public debt ratios in selected OECD countries in 2009 (% of GDP)

	Gross debt ratio	Net debt ratio
Australia	19.2	−3.8
Austria	70.3	37.2
Belgium	101.0	80.7
Czech Republic	42.1	−0.6
Finland	52.6	−63.2
France	86.3	50.6
Germany	76.2	48.3
Greece	119.0	87.0
Iceland	122.7	41.0
Ireland	70.3	27.2
Italy	128.8	101.0
Japan	192.9	108.3
Luxembourg	18.2	−46.1
Norway	49.2	−153.4
Portugal	87.0	57.9
Spain	62.6	34.8
Sweden	51.8	−23.4
UK	72.3	43.5
US	83.0	58.2
Euro area	86.3	53.8
OECD	90.3	51.5

Note: Gross debt ratios are measured by the OECD at market value and may thus differ from other figures quoted in this chapter, which are sometimes measured at face value.

Source: *OECD Economic Outlook* no. 87, April 2010.

and difficult to evaluate. Finally, and most importantly, governments can change the very laws under which they operate, blurring the notion of a government "liability." For example, a reform of a pay-as-you-go pension system often amounts to a legal default on previous implicit liabilities. As a result, there are debates on what government balance sheets should look like, in particular when it comes to *off-balance liabilities** such as pension rights accrued to civil servants or guarantees extended by the government to private undertakings. Table 3.2 shows that accounting for pension provisions led to an increase by 40% of the amount of the US federal debt in 2009. A similar ratio is obtained in the case of France. Off-balance liabilities are crucial to assessing public finance sustainability, an issue to which we shall return in the next section.

Table 3.2
The US federal government balance sheet on 30 September 2009 (US dollars, billion)

Assets		Liabilities	
Cash and other monetary assets	393.2	Federal debt securities	7582.7
Securities and investment	93.1	Other liabilities	6541.1
Other assets	2181.6	of which: Federal employee and veteran benefits payable	5283.7
		Net position	−11 455.9

Source: US Treasury (2009) *Financial Report of the United States Government.*

3.1.2 Lessons from history

A cursory glance at history points to a number of stylized facts.[10] We focus here on five of them:

1. A generalized practice of public deficits has developed in the 1970s.
2. It has resulted in growing public debt levels and, for some countries, in a deterioration of public debt to GDP ratios.
3. Debt ratios reached at the beginning of the twenty-first century were appreciably lower than some of the debt ratios experienced in the past, which could eventually be substantially reduced; the fiscal response to the severe economic crisis that started in 2008, however, resulted in a significant increase in debt ratios in many countries, of a scale unprecedented since the end of World War II.
4. The developments of the 1990s and 2000s reflect very different philosophies concerning the use of fiscal policy.
5. The effects of an active use of fiscal policy, whether toward expansion or contraction, are stable neither in time nor in space.

a) Taxes, expenditures, and deficits

Figure 3.7 shows that the almost-systematic practice of budget deficits in the major industrialized countries dates from the early 1970s. Where did these deficits come from? Figure 3.8 highlights the increasing share of public expenditure in GDP from 1970 on. What is the relation between these two developments?

The long-term trend of rising public expenditure levels owes much to the increase in social insurance expenditures—including in countries such as the

10. For a survey, see for example Masson and Mussa (1995) and, for Europe, Bismut and Jacquet (1997). See also, in the context of the 2008–09 crisis, International Monetary Fund (2009).

US and Japan where these expenditures represent a lower share of GDP than in Europe—and to the rising burden of interest payments on the debt.

Until the 1970s, the rise in public expenditures was paralleled with a regular increase in public receipts. But from the 1970s on, public receipts fell short of spending. In response, and to slow down debt accumulation, governments initially tried to stabilize expenditures as a percentage of GDP. But an additional priority soon surfaced: Tax cuts. Increasingly, supply-side economists and conservative political leaders highlighted the negative consequences of excessive tax rates on the operation of the economy (see chapter 7 for a discussion of the distortionary effects of taxes). In the early 1980s, for Ronald Reagan in the US and for Margaret Thatcher in the UK, cutting taxes became a central economic policy objective.

Against this background the traditional logical sequence—expenditure control, deficit reduction, and eventually tax cuts—was in the US replaced by a new one: Cut taxes first, and then force policymakers and the public to face the trade-off between higher deficits or cuts in expenditures. In political economy terms, the objective was to "starve the beast" and make expenditure cuts unavoidable. The situation was reversed only in the second half of the 1990s in the US, with the combination of faster growth and stricter expenditure control, before the country plunged again into higher deficits in the 2000s following new tax cuts accompanied by an economic downturn. Until the 1990s, mainland Europe stuck to the traditional sequence. But favorable economic conditions at the end of the 1990s allowed reduction of taxes.

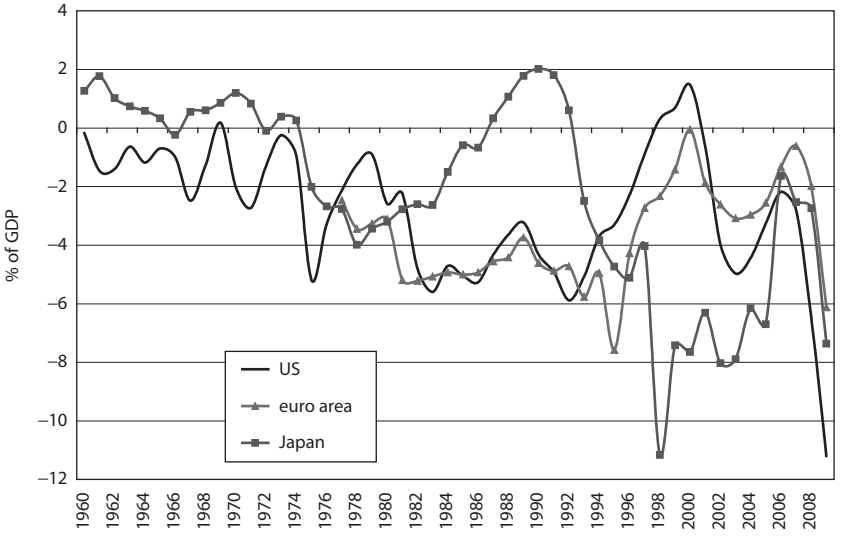


Figure 3.7 General government financial balances in the US, the euro area, and Japan, 1960–2009.
Source: *OECD Economic Outlook* no. 86, November 2009.

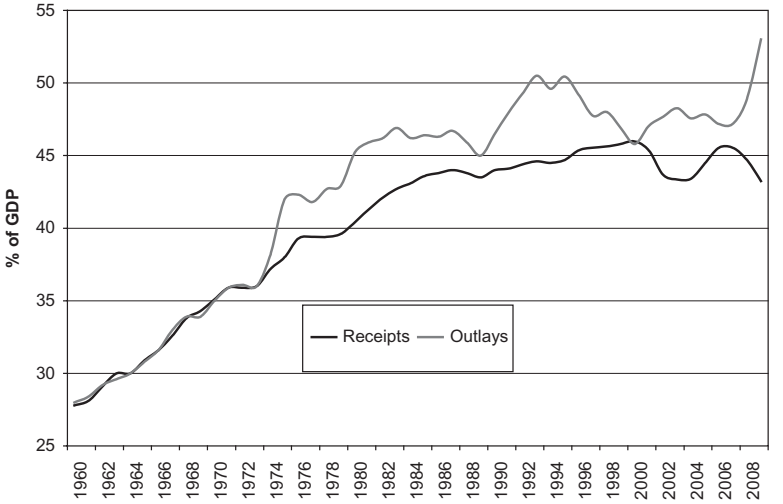


Figure 3.8 Public expenditure and receipts in the OECD countries.
Source: OECD.
Note: There is a break in the series in 1995. Recent data have been adjusted to fit the historical series.

The subsequent increase in structural deficits was at the root of the difficulties of the following years.

b) Debt dynamics

Until the end of the 1970s, thanks to a high nominal growth compared to the level of interest rates (see box 3.2), it was possible to sustain primary deficits without increasing the debt burden. The situation reversed in the 1980s, when a higher real interest rate combined with an economic slowdown to accelerate debt accumulation. In Europe, the public-debt-to-GDP ratio of Belgium and Italy jumped in a decade from 50% or 60% to more than 100% (see figure 3.9). Ireland also witnessed a marked increase of its debt ratio that the economic growth of the 1990s then made it possible to control. France, for its part, had a small public debt at the end of the 1970s, with a debt-to-GDP ratio of about 20%. However, the progression of this ratio has been almost continuous since. Outside Europe, Japan has experienced a spectacular rise, with a debt ratio multiplied by more than 17 between 1969 (10%) and 2007 (172%). During the 2007–09 crisis, all these countries experienced a jump in their public debt ratios as a consequence of crisis-related expenditures, lower revenues, and the cost of recapitalizing ailing banks. Between 2007 and 2009 the debt ratios increased by 22% of GDP in the US, 24% of GDP in the UK, 11% of GDP in the euro area, and as much as 64% of GDP in Iceland.

A high debt ratio leaves public finance more vulnerable to interest-rate rises, with the risk of having to devote increasingly important resources to servicing

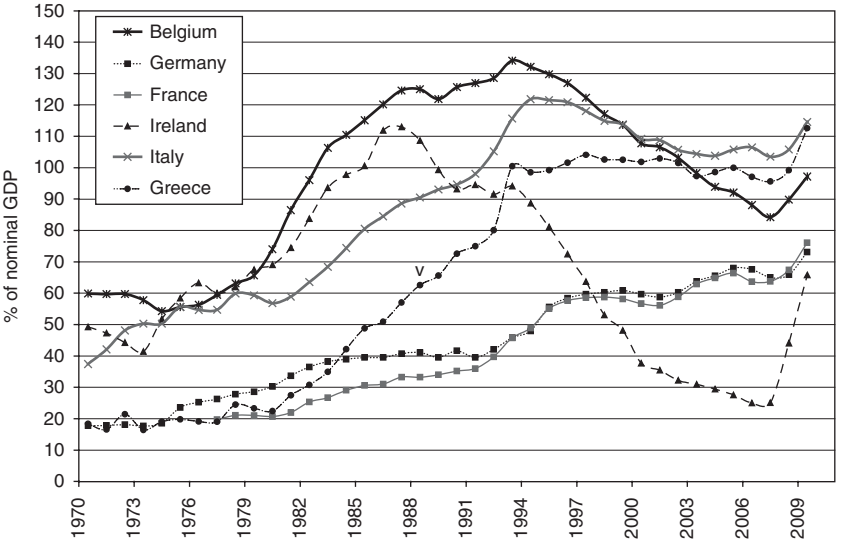


Figure 3.9 Public debt developments in selected European countries.
Source: European Commission (AMECO database and ECFIN November 2009 forecasts; data for Germany before 1990 are for West Germany).

the debt and of having no other option for financing a given expenditure program than to raise taxes, which is likely to distort the economy and to be politically and socially costly. In Belgium, for example, interest payments on the public debt reached 11.4% of the GDP in 1986: More than 20% of tax income thus had to be devoted to them. A high debt level can also instill fears regarding the capacity of the government to service it. It also can, indirectly, weaken the financial system that holds the largest share of public debt.

c) The debt ratio in a historical perspective

Observation of the past, however, allows us to put the rise of public debt ratios in the late 2000s into perspective. Wars formerly resulted in public debt increases up to levels (in terms of percentage of the GDP) sometimes much higher than those observed in the late 2000s. For example, the public debt in the UK exceeded 100% of GDP during more than a century; it reached a record level of almost 300% of GDP in 1821 after the Napoleonic wars, before stabilizing at less than 100% after 1860 (Buiter, 1985); as in other industrialized countries, the debt ratio then underwent a new explosion in the first half of the twentieth century in the aftermath of the two world wars (cf. figure 3.10).

History demonstrates that economies sometimes reach considerable public debt levels but still adjust back to normal. However, this experience does not imply that the cost of adjustment to an excessive debt burden is negligible. For example, one of the causes of the French Revolution was the almost bankrupt state of public finance in the last quarter of the eighteenth century (the French

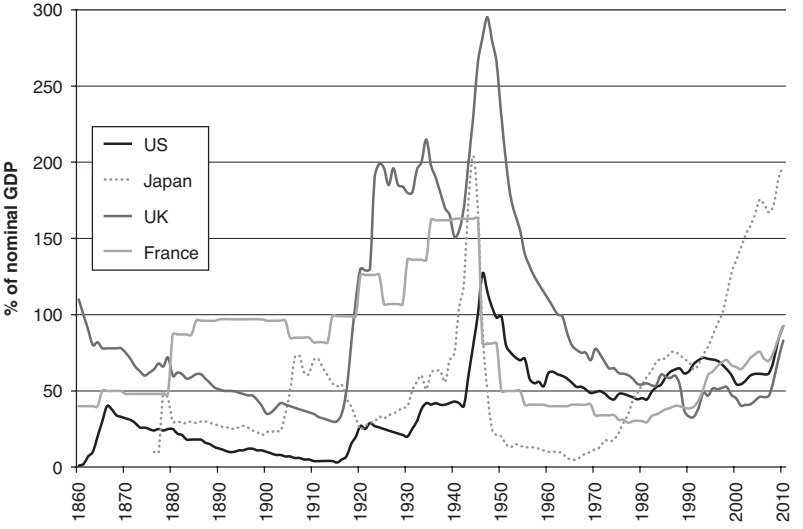


Figure 3.10 Long-term public debt developments in selected countries.
Source: Masson and Mussa (1995), and *OECD Economic Outlook* no. 86, November 2009.
Note: Data for France are five-year averages until 1960.

Kingdom was declared bankrupt on 16 August 1788) and the double challenge, because of existing privileges, of checking public expenditure and of widening the tax base.

Several factors, however, limit the scope of this analogy with history. First of all, debt in peacetime has hardly ever experienced a rise similar to that of the last quarter of the twentieth century. The temporary character of expenditures that were first related to wars and then to reconstruction can explain why, once wars were over, financing needs more or less reverted to their previous level and the debt overhang could be cured. In peacetime, debt reflects a more structural and sustained, rather than temporary, financing need: The rise in public expenditures over the twentieth century, common to all industrialized countries, primarily concerns social insurance, spending on which underwent a particularly marked increase in European countries. During the 1990s, Japan tried to fight its long-lasting crisis with multiple fiscal stimulation packages that led to a dramatic increase in public debt. After 2008, all industrial countries followed suit in the aftermath of the world crisis (see figure 3.10). Looking forward, population aging suggests that public spending in industrialized countries on retirement and health care is bound to increase substantially. For instance, the US Congressional Budget Office calculated in 2003 that based on then-scheduled benefits, Medicare, Medicaid[11] and Social

11. Medicare and Medicaid are programs targeting access to healthcare by low-income households and retirees in the US.

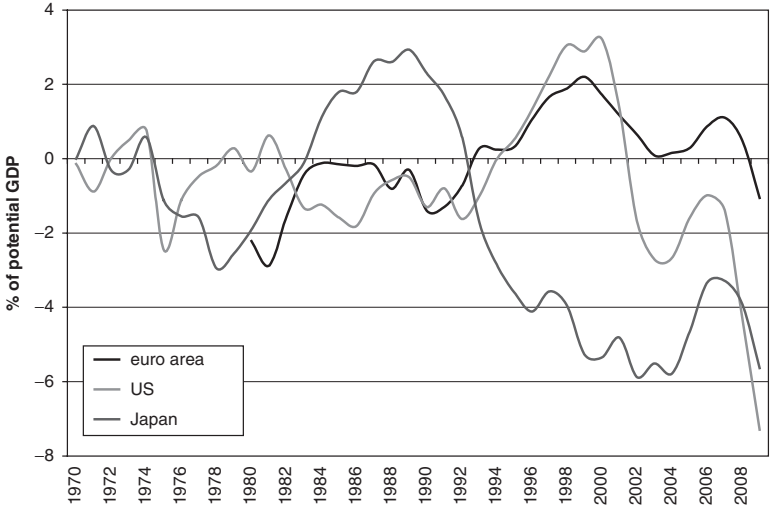


Figure 3.11 Underlying primary balances: US, euro area, Japan.
Source: *OECD Economic Outlook* no. 86, November 2009).
Note: Underlying primary balances from 1991 on; primary structural balances before 1991.

Security expenditures would rise by 9.4 percentage points of GDP from 2003 to 2050 (see OECD, 2005).[12] As already discussed, pay-as-you-go pension schemes also represent large off-balance liabilities that lead to a more alarmist diagnosis. Highly indebted governments are ill-prepared to face the need for future surges in social expenditures.

d) Different philosophies and policies from country to country

The use of fiscal policy has varied markedly across industrial countries since the 1970s. Figure 3.11 shows how fiscal policy has been very actively used in the US and even more in Japan, as opposed to the euro area's apparent prudence.

The United States: Fiscal activism Activism in the use of fiscal policy instruments emerged in the US in the 1960s. Walter Heller, who was chairman of President Kennedy's *Council of Economic Advisers*, thus summarized the corresponding philosophy:

> Economy has come of age in the 1960s. Two Presidents have recognized and drawn on modern economics as a source of national strength and

12. The 2009 health care reform is expected to slow down these costs from an annualized 6.4% growth rate before the reform to a 5.6% rate after it, over the 2016–19 period, after an initial increase by $10bn (1% of total cost) due to extended coverage of these programs (see Executive Office of the President and Council of Economic advisers, 2009).

> Presidential power. Their willingness to use, for the first time, the full range of modern economic tools underlies the unbroken US expansion since early 1961—an expansion that in its first five years created over seven million new jobs, doubled profits, increased the nation's real output by a third, and closed the $50-billion gap between actual and potential production that plagued the American economy in 1961.
>
> W. Heller (1966), p. 1

In the 1970s, the US responded to the first oil shock with fiscal expansion (Japan and several European countries did the same). However, this experience with Keynesian expansion was shorter-lived in the US than in Europe, and by the late 1970s the underlying primary deficit had been eliminated (figure 3.11). A major turning point took place in the early 1980s when Ronald Reagan's administration introduced both significant tax cuts and additional military spending. Although this program was supposed to have been inspired by "supply-side economics" (cf. chapter 1), the short-term impact was in fact that of a Keynesian expansion. The structural fiscal balance deteriorated substantially and remained in deficit until the beginning of the 1990s. As the current account deteriorated simultaneously, the 1980s were marred with *twin deficits** (of the budget and the current account). Beginning in 1993, the Clinton administration embarked on a fiscal adjustment policy based on a strict control on spending, and from 1994 until 2000, the US general government budget recorded a structural primary surplus. Public debt was reduced both as a percentage of GDP and in dollar terms and there was even consideration of its prospective extinction before it started increasing again at a rapid pace in the 2000s (figures 3.6 and 3.10).

The 2001–02 cyclical downturn and George W. Bush's election led to a dramatic policy reversal that in two years transformed a primary structural surplus of 3.5% of GDP into a deficit of more than 2% of GDP (cf. figure 3.11), which, in the context of the low phase of the economic cycle, resulted in a total fiscal deficit of about 5% of GDP in 2003. The relevance of this policy was vigorously questioned, in its macroeconomic as well as its redistributive aspects.[13] The related debates did not focus, as in the euro area, on comparing the deficit or debt ratios with any threshold, but rather on the economic doctrine that underlay the administration's choices and the economic relevance of these choices. The 2007 subprime crisis, followed by the most severe economic crisis since the Great Depression, paved the way for an aggressive fiscal response and a marked deterioration of the underlying primary deficit to record levels (for the US), exceeding 7% of GDP in 2009.

Beyond differences of doctrines and practices between various US administrations, the US fiscal policy is marked by the permanence of its discretionary approach. Reliance on it during recessions to accelerate the recovery is not

13. See Stiglitz (2003), or Paul Krugman's regular columns in the *New York Times* in 2002 and 2003 and the criticism formulated by the IMF in its annual report of 2003 on the US.

questioned in policy circles, even though, at least before the severe economic recession of 2008–09, the US economic profession had largely turned against the proactive use of discretionary fiscal policy (see for instance Taylor, 2000, 2009; Feldstein, 2002). The idea of fixing numerical targets for the federal fiscal balance or public debt ratio is fundamentally alien to the US conception, even though most individual US states operate under some form of balanced-budget rule.

Japan: Low-return stimulus efforts Japan used fiscal policy massively in the 1990s in an attempt to restore growth after it had vanished in the aftermath of the bursting of the speculative bubble of the end of the 1980s. Twelve expansionary plans involving, in particular, public investment programs were announced between August 1992 and February 2002, some of which accounted for more than 2% of GDP (OECD, 2002). As a consequence the structural balance deteriorated by more than eight percentage points of GDP between 1989 and 2003 (cf. figure 3.11). The gross public debt, which was negligible in the 1960s, reached more than 160% of GDP in 2003 and 180% in 2007, the highest level within the OECD. Some fears even emerged about the capacity of Japan to honor its debt, as illustrated by the deterioration of the grade given to Japan by rating agencies.

The effectiveness of this historically unprecedented stimulus is generally considered to have been weak: Expenditures by the government did not have a significant stimulating effect on private behavior and Japanese growth did not recover durably.[14] This led to a gradual recognition that the major obstacles to recovery were the protracted deleveraging process going on in the private sector and the impaired assets of the banking sector, which constrained banks' ability to lend and therefore hampered the transmission of monetary policy and acted as a drag on private spending. Furthermore, Japan was facing structural problems, such as the need to adapt to globalization and dismal demographic prospects.

In 2008–09, Japan again decided on a new series of fiscal expansions in response to the economic crisis. Five successive fiscal packages were introduced from August 2008 to spring 2009 in order to tame the recession and react to a dramatic shrinking of the Japanese economy (−12.1% in annualized terms over the last quarter of 2008). As a result, the underlying primary deficit worsened after a significant improvement from 2004 to 2007 (see figure 3.11) and the gross public debt came close to 200% of GDP.

The euro area: Constrained muddling through Whatever the indicator, Europe in the 1980s was characterized by a severe deterioration of its fiscal situation, which explains the later focus on public finance adjustment. The situation improved in the second half of the 1990s, particularly in the run-up

14. A study by Kuttner and Posen (2002) claims, however, that Japan's fiscal expansions had a significant impact, especially when delivered through tax cuts.

to monetary union, because a deficit criterion was used to judge the capacity of the member countries to take part in the euro (see below). However, after the initial effort, fatigue set in: The reduction of fiscal deficits continued after 1997, the year in which the entry examination for admission into the euro area in 1999 took place; but this was mainly due to stronger growth and to a decline in long-term interest rates in countries whose qualification was dubious. On the whole, the underlying fiscal balance improved in the 1990s but fluctuated around a 1% surplus in the first half of the 2000s, in spite of the participating countries commitment to go beyond. The economic downturn in the early 2000s led member countries to adopt very different fiscal policies: Finland and Ireland used their margins of maneuver to implement a counter-cyclical policy; some like Belgium, Austria, or Spain gave priority to the pursuit of adjustment; others, like France and Germany, refused to follow a restrictive fiscal policy in a period of downturn, while they had no formal margin for pursuing stimulus on a large scale. Fiscal adjustment was deliberately amplified after 2003, but Italy moved in the opposite direction until 2006 and France's efforts were cosmetic. After 2008, the economic recession brought fiscal adjustment to a stop, as euro area governments decided to let automatic stabilizers play their role and also responded with expansionary fiscal policies.

On the whole, as illustrated in figure 3.12, notwithstanding a counter-cyclical contraction in 2006 and 2007 and the counter-cyclical expansion of 2008–09, the euro area tends to have practiced a *pro-cyclical policy**—i.e., one that accentuated aggregate demand fluctuations—or a neutral one, while the US used fiscal policy in a more *counter-cyclical** way—i.e., with a view to dampening fluctuations in aggregate demand.

e) Unstable outcomes

Did proactive fiscal policies have the expected impact on the economic activity? Empirical observations do not lead to unambiguous conclusions. Some fiscal expansions clearly boosted activity: The US stimulus episodes of the 1980s and the 2000s fall into this category. Other episodes, however, suggest that this effect is not systematic. For example, between 1982 and 1986 a fiscal contraction of almost 10% of GDP in Denmark was accompanied by an economic recovery and by vigorous growth (see figure 3.13). In fact, the literature suggests that the economy does not seem to respond in a systematically Keynesian manner to large-scale fiscal policies. Fiscal expansion can be ineffective while fiscal contraction can have an expansionary impact (so-called *anti-Keynesian effects**). In some cases, such as in Sweden in the early 1990s and Japan in the 1990s, this can be explained by the simultaneous effect of banking crises. However, there is also evidence that an unsustainable fiscal stimulus may lead to precautionary saving by private agents in the expectation of ensuing adjustment. In contrast, at the end of 2008, there was a sizeable consensus among economists that fiscal packages would help limit the effects

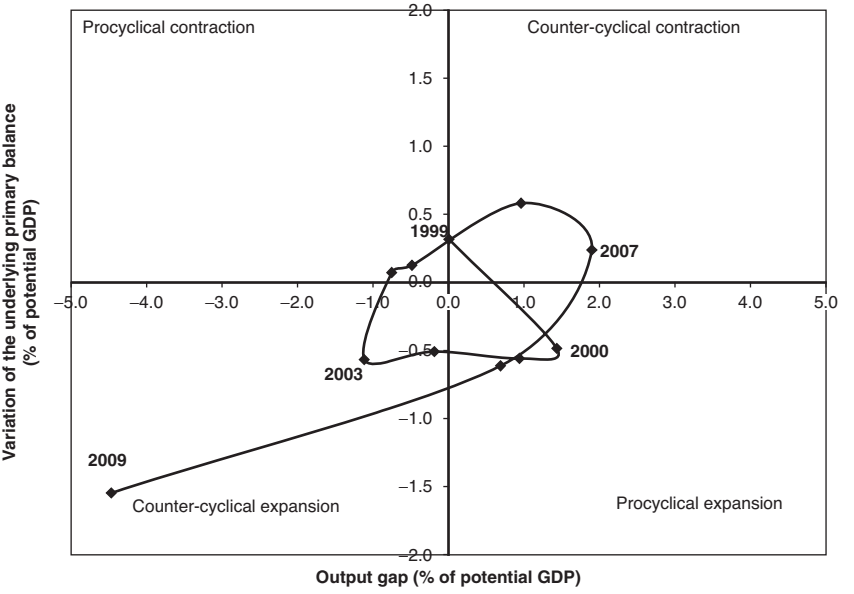


Figure 3.12 Fiscal policy stance in the euro area, 1999–2009.
Source: *OECD Economic Outlook* no. 86, November 2009.

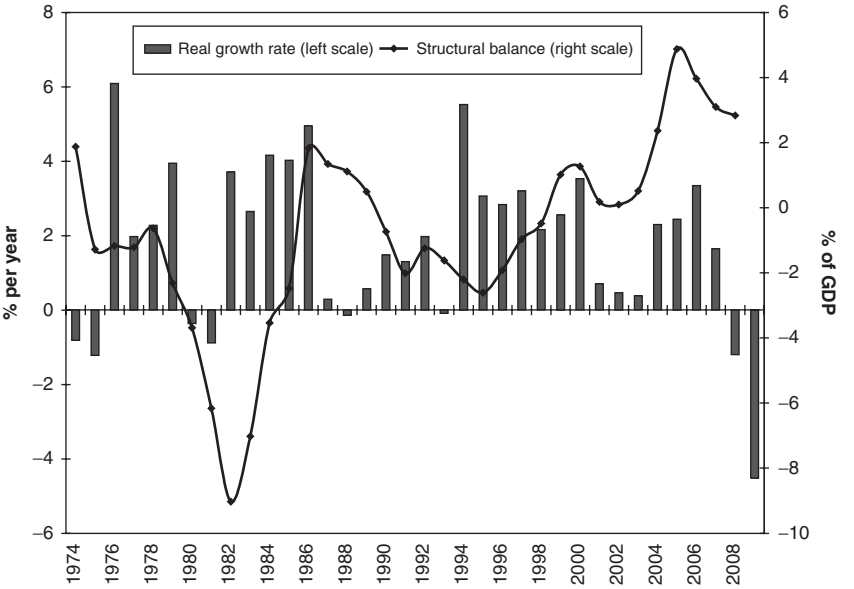


Figure 3.13 Fiscal policy and GDP growth in Denmark.
Source: *OECD Economic Outlook* no. 86, November 2009.

of the crisis, in a context where, due to the state of the banking sector, credit constraints were more acute than in normal times.

3.2 Theories

John Maynard Keynes's *General Theory of Employment, Interest and Money* (1936) has provided, since its publication, the conceptual framework for the use of fiscal policy to influence the level of aggregate demand. Whereas the classical theory was primarily concerned about public finance solvency, in other words about the debt *stock*, Keynes's analyses focused on the role of *flows* of public receipts and expenditures in the determination of the aggregate macroeconomic equilibrium. By definition, however, debt results from the accumulation of deficits. Yet, this obvious fact was consistently ignored in the first three decades after World War II. It was only in reaction to an excessive reliance on fiscal policy in the 1970s, to the associated permanent deficits and to the resulting increase in public debt ratios that debt-related concerns gradually came to the fore. In response to these concerns, economists developed models to represent public debt dynamics and their effects on the economy.

In this section, we first briefly sketch the Keynesian theory and the main criticisms of it. We then examine the dynamics and sustainability of public debt. We finally present more-comprehensive approaches that combine in a single model issues of debt sustainability and fiscal policy effectiveness.

3.2.1 Demand-side effects: Keynes and his critics

a) The Keynesian analysis

As indicated in chapter 1, the standard Keynesian approach starts from the assumption of price rigidity or at least stickiness in the short term. This implies that prices do not adjust immediately to ensure macroeconomic balance. In other words, the supply of goods and services is elastic and macroeconomic balance—output and employment—is determined by the level of aggregate demand. When aggregate demand is insufficient, this results in the underemployment of production factors in the economy. A fundamental role of macroeconomic policy—be it fiscal or monetary—is to ensure that the level of aggregate demand is such that the economy remains at, or close to a level corresponding to full employment.

In the elementary model, nominal rigidity is simply postulated, or it is regarded as a fact of life resulting from the existence of contracts specified in nominal terms. Since the 1980s, however, the so-called "New-Keynesian" economists have developed micro-founded models of nominal rigidities relying on optimizing behavior by individual agents (see chapter 4). Another assumption is that households are somewhat myopic so that consumption depends on current income (a more sophisticated explanation is

that they do not have access to financial markets, and therefore cannot smooth their consumption levels over time).

Under these conditions, macroeconomic equilibrium does not result from price movements; rather, it is determined by the level of aggregate demand. An exogenous variation in aggregate demand (a demand shock) results in a proportional variation in the level of output. The ratio between output variation and the initial exogenous variation of aggregate demand is called the *Keynesian multiplier** (box 3.4).

Box 3.4 A Primer on the Keynesian Multiplier

Suppose that household consumption C is a linear function of current income Y:

$$C = aY + b,\ a, b > 0 \qquad \text{(B3.4.1)}$$

The parameter a is the *marginal propensity to consume** (meaning that out of one additional dollar or euro of disposable income, households spend a and they save $(1 - a)$). Let us assume that $a = 0.8$, so that households consume 80% of any additional unit of income.

Suppose that supply is perfectly elastic, so that output adjusts to the level of aggregate demand at constant prices. The product market equilibrium is written as:

$$Y = C + \bar{I} + \bar{G} \qquad \text{(B3.4.2)}$$

where $\bar{I}$ is aggregate investment and $\bar{G}$ is government demand. Both are assumed to be exogenous.

Suppose the government increases public spending by one unit (and assume for the time being that there is no tax increase). This will initially lift output, and thus income distributed to households, by one euro. Out of this additional unit, 80 cents will be consumed and will lift output—thus disposable income—further. At the end of the process, the total increase in output is:

$$\begin{aligned} 1 + a + a^2 + a^3 + \ldots &= 1 + 0.8 + 0.8^2 + 0.8^3 + \ldots \\ &= 1/(1 - a) = 1/(1 - 0.8) = 5 \text{ euros} \end{aligned}$$

Hence, we have:

$$\Delta Y = \frac{\Delta \bar{G}}{1 - a} \qquad \text{(B3.4.3)}$$

In this example the multiplier is very large and fiscal policy is therefore extremely powerful. There are however many factors that may lower

the multiplier:

1. Not all the additional income accrues to consumers. A fraction may be retained by firms in the form of retained earnings. Even disregarding this factor, another fraction is necessarily taxed away by the government. So equation (B3.4.1) needs to be rewritten $C = a(1 - t)Y + b$ where t is the tax rate and the multiplier becomes $1/[1 - a(1 - t)]$
2. In an open economy, an additional euro of disposable income leads households to consume more of both domestic and imported products and firms to import more intermediate goods. Assuming that the marginal propensity to import is m (meaning that an additional euro of income will lead to m euros of imports), the Keynesian multiplier becomes $1/[1 - a(1 - t) + m]$.
3. The assumption of complete price rigidity is extreme. If prices adjust upward, part of the increase in demand does not result in an increase of the *volume* of products consumed but in an increase in their price. This especially applies over time, as prices adjust gradually.
4. The central bank may respond to an increased demand for products with a less accommodative monetary policy and engineer a rise in the interest rate. In this case investment demand (from firms) declines because firms compare the yield of investment projects with the financing cost or with the return to financial investments. A *crowding-out effect** appears: Part of the increase in public demand results in lower private investment by firms (due to the interest-rate increase, private investment is crowded out by public demand).

All these factors weaken the impact of fiscal expansion on aggregate demand and income.

The Keynesian assumptions can be represented within the "aggregate supply, aggregate demand" (AS–AD) model presented in chapter 1. The price stickiness assumption implies that the aggregate supply (AS) curve is upward-sloping but not vertical in the short run. In the elementary model, the slope of AS is low, so that supply is highly responsive to price movements. Production can therefore be increased or decreased without a major impact on prices. The aggregate demand curve is downward-sloping due to the negative impact of inflation on demand for goods and services, either through a wealth effect or through the impact of an endogenous rise of the interest rate. A fiscal expansion (through a rise in public spending or a cut in taxes) results in the demand curve moving to the right: Production increases at any given price level. If the slope of the supply curve is low, this does not have a major impact

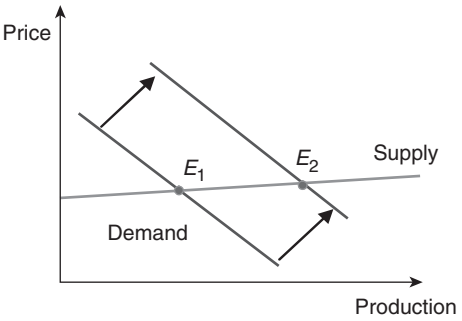


Figure 3.14 Effect of a Keynesian expansionary fiscal policy.

on the price level and the adjustment takes place through a variation in the output level (movement of E1 to E2 in Figure 3.14).

Here we have simply postulated the AD curve that summarizes the demand side of the economy. It can be derived from the *IS–LM model** introduced in the late 1930s on the basis of Keynes's *General Theory*.[15] This model, which has been widely used ever since to represent the fixed-price Keynesian model, consists of two curves that relate output and the interest rate: The IS curve describes the product market equilibrium and the LM curve the money market equilibrium, both *at a given price*:

- The IS curve represents the combination of output and interest rate that results in a product market equilibrium. It is downward-sloping since a higher interest rate results in a lower demand for products;
- For a given money supply, the LM curve shows the combination of output and interest rate that results in a money market equilibrium. With a fixed money supply, the positive relationship between output and the interest rate relies on the demand for money, which is supposed to be an increasing function of output (as output grows, more money is needed for transactions) and a decreasing function of the interest rate (as the interest rate grows, private agents prefer to hold interest-bearing assets rather than cash).[16]

The solution of the IS–LM model shows equilibrium output and interest rate for a given price level. As price grows, the demand for products declines

15. The IS–LM model was introduced in 1937 by Sir John Hicks (Hicks, 1937) and popularized by Alvin Hansen (for example, Hansen 1953). For a presentation of the model, see Blanchard (2005) or Mankiw (2007).

16. Modern analysis of interest rate formation no longer starts from a given money supply. Rather, the short-term interest rate is supposed to be set by the central bank in response to economic developments, in order to ensure macroeconomic stability in the medium run. As a result (this is further developed in chapter 4), the interest rate becomes an increasing function of the demand for goods and services, which is analogous to the formulation of the LM curve.

(because the real value of nominal balances diminishes, making consumers poorer), the IS curve shifts downwards and the equilibrium level of output is also lower. This implies that the AD curve is downward-sloping.

A fiscal expansion is represented in the model as a shift to the right of the IS curve. For any given price level, the fiscal expansion results in a higher output and interest rate, therefore in a shift to the right of the AD curve as in figure 3.14.

Because monetary and fiscal policies are to a large extent substitutable, the Keynesian approach naturally leads to thinking in terms of *policy-mix**, i.e., of combination of them. In particular, in this framework fiscal policy is more effective when it is supported by monetary policy. At the limit, a perfectly *accommodative monetary policy** that does not lead to increasing the interest rate in response to a fiscal expansion results in a maximum multiplier effect. When the central bank is independent, however, it may choose not to accommodate the effects of fiscal policy if it perceives it as potentially inflationary. In the representation in figure 3.14, if the supply curve is steep monetary policy is likely to react. Generally speaking, fiscal policy cannot be studied in isolation from monetary policy.

The Keynesian approach can easily be extended to the open economy, in particular within the *Mundell–Fleming model**, developed independently by Robert Mundell and Marcus Fleming in the early 1960s.[17] This open-economy extension of the IS–LM model introduces the exchange-rate regimes as a key determinant of the Keynesian multiplier. In a flexible exchange-rate regime, the fiscal multiplier is lowered—even nullified if capital is perfectly mobile across countries—by the appreciation of the exchange rate that follows a fiscal expansion. Conversely, the multiplier is larger in a fixed exchange-rate regime because there is little crowding out (see box 3.5).[18]

Given its simplicity, the Mundell–Fleming model remains a widely used reference by international-economy practitioners.

Box 3.5 The Mundell–Fleming Model

The canonical Mundell–Fleming model studies policy effectiveness in a small country under *perfect capital mobility* (and under the Keynesian assumption of underemployment of resources).

Perfect capital mobility implies that the interest rate cannot deviate from the world interest rate (otherwise capital would flow in or out in search of yield). This is represented by the horizontal interest rate arbitrage condition schedule. At the same time, the internal equilibrium is represented by the IS and LM curves, representing respectively product market and money market equilibrium.

17. See Mundell (1968), Fleming (1962).
18. Exchange-rate regimes are defined in chapter 5.

The open-economy equilibrium seems to be overdetermined since it results from the intersection of three different curves. To see how the model works, one needs to distinguish the cases of floating and fixed exchange rates.

Consider first the case of a *floating* exchange-rate regime. Assume that the central bank keeps money supply constant and that the exchange rate is market-determined. A fiscal expansion leads to an increase in output and income, and thereby to an increase in money demand. With constant money supply, there is a rising pressure on the interest rate. This leads to capital inflows that cause an exchange rate appreciation and a loss of export competitiveness. The IS curve thus shifts to the left as the demand for the country's products diminishes. Since the open-economy equilibrium is determined by the intersection of LM and the international interest rate arbitrage condition, the only solution is that the exchange rate appreciates up to the point where aggregate demand returns to its original level before the expansion (cf. figure B3.5.1). The IS curve plays no role in the determination of the equilibrium and fiscal policy has no impact on output. Public demand here crowds out not the residents' investment (in a small country under perfect capital mobility, the interest rate remains fixed at the world level *ex post*), but the nonresidents' net demand for the country's exports.

This effect can be illustrated by the strong fiscal expansion carried out by Ronald Reagan's administration in 1981, at a time when monetary

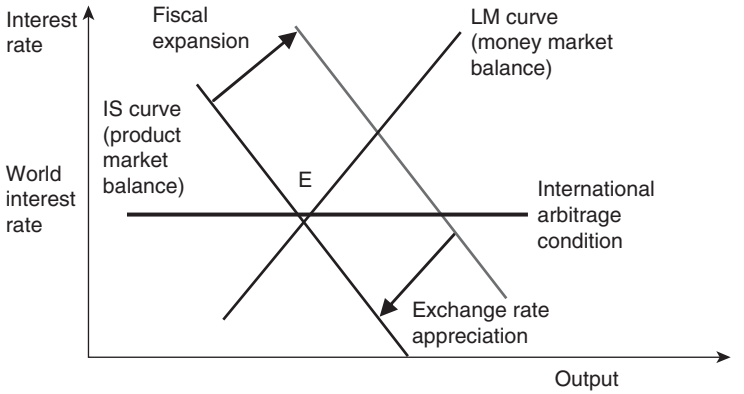


Figure B3.5.1 Fiscal expansion under flexible exchange rates and perfect capital mobility.
Reading: The fiscal expansion moves the IS curve to the right because aggregate demand is higher for a given interest rate. However this raises money demand for transaction purposes. The upward pressure on the interest rate generates capital inflows. The exchange rate appreciates and the consequent loss in competitiveness brings the IS curve back to its initial position.

policy was geared toward controlling the money supply and therefore fully nonaccommodating. The fiscal expansion led to a sharp increase in the interest rates, a marked appreciation of the US dollar, and a very significant deterioration in the current account balance.

Now suppose that the exchange rate is *fixed*, meaning that the central bank intervenes on the foreign exchange market through buying and selling foreign currency. The capital inflows consecutive to a fiscal expansion result in the central bank selling the domestic currency for the foreign one, and thereby in an accumulation of foreign exchange reserves by the central bank. This increases the money supply and makes the LM curve move to the right. This endogenous monetary expansion leads to a positive fiscal multiplier (cf. figure B3.5.2).

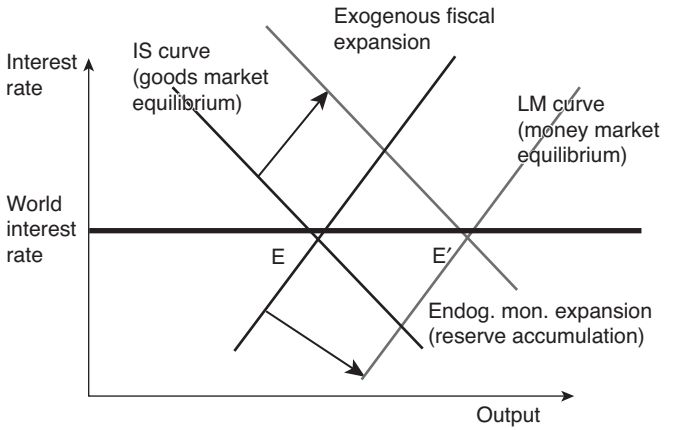


Figure B3.5.2 Fiscal expansion under fixed exchange rates and perfect capital mobility.
Reading: The fiscal expansion moves the IS curve to the right because aggregate demand is higher for a given interest rate. The upward pressure on the interest rate generates reserve accumulation by the central bank (to prevent exchange-rate appreciation). The rise in money supply moves the LM curve to the right. *Ex post*, the interest rate and exchange rate are unchanged and the fiscal multiplier is positive as the equilibrium moves from E to E'.

When capital is *not mobile*, results are reversed: Under a *floating* exchange regime, the deterioration of the current account balance induced by fiscal expansion (because of the increase in import demand) leads to an exchange-rate depreciation and to an improvement of export competitiveness, strengthening the initial demand impact of the fiscal expansion. Under *fixed* exchange rates, the current account deterioration results in a reserve loss and in a monetary contraction that counters the

initial expansion. Ultimately, the current account must balance—which means that output is determined by the external constraint.

Differences in capital mobility and exchange-rate regimes thus explain why similar fiscal policies can have contrasted effects on output. For example, at the same time that Ronald Reagan's expansionary policies pushed the dollar upward, the fiscal expansion undertaken in France by the socialist government under the new President François Mitterrand created downward pressures on the French franc. The explanation of this difference is that while the US had already liberalized foreign exchange, France had not.

The main results of the model are summarized in table 3.3. A monetary union behaves as a whole like a flexible exchange-rate regime in relation to the rest of the world. If capital is fully mobile, fiscal policy is relatively ineffective. However, for a given member of the zone, fiscal policy is effective because crowding-out effects are diluted within the zone. Hence, the Mundell–Fleming model suggests the use of fiscal policy by individual member states that may be hit by asymmetric shocks[19] but less so as a collective response to a symmetric shock, because in the latter case the exchange-rate adjustment would partially offset the stabilizing effect of fiscal policy.[20]

Another way to present the same results puts the emphasis on the international spill-over effects of fiscal policy. In a floating exchange-rate regime and under capital mobility, the effect of a fiscal expansion at home increases the demand for foreign goods (the appreciation of the exchange rate reinforcing the direct, demand effect). Concerted fiscal expansion therefore restores domestic effectiveness, because if all countries embark on a fiscal expansion, exchange rates do not move. In fact, the effects of a concerted fiscal expansion are identical under fixed or floating exchange rates and (other things being equal) the multiplier lies in-between the high fixed exchange-rate multiplier and the low floating exchange-rate multiplier.

Table 3.3
Short-term effectiveness of fiscal policy in an open economy

	High capital mobility	Low capital mobility
Floating exchange rates	Ineffective or not very effective	**Effective**
Fixed exchange rates	**Effective**	Not very effective

19. Asymmetric shocks differ from one country to another, as opposed to symmetric shocks, see chapter 1.
20. Since the euro area is not a small economy, the crowding out by the exchange rate is less than perfect in this case.

b) The neoclassical critique

The neoclassical critique of the multiplier rests on three separate arguments:

- *Full financial crowding-out*: After a fiscal expansion, the deterioration of the public balance causes a rise in the interest rate which depresses private demand (crowding-out effect). In the AS–AD model, the demand curve does not move (or moves little) in the event of a fiscal shock: Total demand is not affected by a rise of public demand, but its composition is modified by the substitution of public for private demand.
- *Supply rigidity*: The relative price adjustment is sufficiently rapid so that the goods–market equilibrium is determined by supply. In the AS–AD model, the demand curve moves toward the right but the supply curve is very steep and almost vertical: Producers agree to slightly increase supply only if prices increase a lot. Private demand is penalized *ex post* by the rise in prices (cf. figure 3.15).

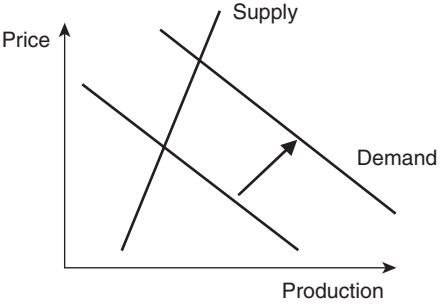


Figure 3.15 Effect of an expansionary fiscal policy with inelastic supply.

- *Ricardian equivalence**: Even if the supply of goods and services is elastic, rational households will respond to an increase in public demand (or a cut in taxes) by restricting their consumption, because they expect today's deficit to translate into higher future taxes and they prepare for it by increasing their savings rate. If their discount rate is equal to the interest rate on public debt, the present value of the expected future taxes will be exactly equal to the cut in current taxes. Accordingly, households' wealth does not change and the tax cut does not have any effect on the activity (see box 3.6). In the case of an increase in public demand, they will also cut their private consumption by the same amount, with the result that aggregate demand does not change. Again, there is full crowding-out, but this time due to households' expectations. The interest rate does not move.

Regarding the first argument, a rise in the interest rate unquestionably penalizes private investment, which affects demand and, in the long run,

harms capital accumulation. This is illustrated in table 3.4 by the fiscal multiplier being lower when monetary policy is allowed to react endogenously. However, the relevance of financial crowding-out has been greatly reduced by international capital mobility, which limits the possibility for the long-term domestic interest rate to differ from the world interest rate, except when the very poor state of public finances induces a *risk premium** that compensates asset holders for the risk that the debt is not refunded (see chapter 4).

The second argument raises an empirical question: What is the slope of the supply curve? Available estimates suggest that it is upward-sloping in the short run, which leaves room for fiscal policy effectiveness. All depends, in fact, on the selected time-horizon: Within a period of a few months or quarters, prices are rigid; within a few years, they adjust. Fiscal policy effectiveness (like that of monetary policy, as discussed in chapter 4) is therefore limited in time. This is confirmed by table 3.4, which shows that the multiplier is close to zero after one year.

Finally, the third argument cannot be invoked simultaneously with the first one, since they are contradictory: The first assumes that the public deficit creates a savings shortage which pushes the interest rates upward, while the third one stipulates a domestic rise in private savings in response to public dissaving. The latter argument is attributed to David Ricardo (1817), although one can find earlier insights, such as this diatribe of Mirabeau against former Finance Minister Necker recommending that King Louis XVI not yield to the temptation of getting into debt:

> Your Director of Finance, your Majesty, misleads you. As soon as the State borrows amounts such that its current income cannot even allow to pay the interest, taxation results, whether it is explicitly declared or not. One day will therefore come when a tax has to be introduced in order to collect resources needed to meet the commitments implied by today's borrowing.
>
> Mirabeau (1787), p. 29 (translated by the authors)

The argument was re-introduced in a formal shape by Robert Barro (1974), who showed that infinitely-lived individuals (or, equivalently, altruistic, finitely lived individuals who bequeath their wealth to the next generation) would fully integrate in their current savings decisions the future tax increases needed to repay debt. This argument is regularly invoked to deny the effectiveness of fiscal policies. There is little doubt that contemporary industrialized-country households, who have access to sophisticated financial markets and worry about their future pensions, are more "Ricardian" than those of the 1960s, who had more difficulty borrowing, and had to base their current consumption on their current income. However, full Ricardian equivalence rests on very strong hypotheses (incidentally, Ricardo himself did not believe it to hold):

- *Rational expectations* (cf. chapter 2). Households need to "see through" the effect of the short-term fiscal expansion and anticipate future taxes.

Table 3.4
Impact on GDP of a 1%-of-GDP increase in public consumption during one year (in %). Simulation results using four macroeconometric models

	Assumptions on interest rates	Short run (≤ 1 year)				Long run (> 1 year)			
		Germany	France	UK	US	Germany	France	UK	US
QUEST (European Commission)	Constant interest rate	0.9	0.9	1.0	na	0.0	0.0	0.0	na
	Price level target	0.6	0.8	0.5	na	0.0	0.0	0.0	na
NIGEM (NIESR)	Constant interest rate during one year, then inflation target	1.0	0.8	0.6	na	0.0	0.0	−0.1	na
MULTIMOD (IMF)	Constant interest rate during one year, then inflation target	1.3	1.3	na	1.1	−0.2	−0.2	na	−0.6
INTERLINK (OECD)	Constant interest rate and exchange rate	1.5	0.8	na	1.1	−0.3	0.2	na	0.1

na: Not available.

Source: Hemming et al. (2002).

- *Unproductive public spending.* Fiscal expansion is supposed to have no positive effect on supply, which is unrealistic: Some public expenditures, notably in research, education or public infrastructure, are likely to lift individuals' future incomes because their social return is higher than the interest rate.
- A *perfect functioning of the credit market* (no *liquidity constraints**): In order for households to be indifferent to a change in current taxes (in exchange for future taxes), they must be able to borrow today against lower future taxes, or, on the contrary, to save in preparation for a future tax rise.
- *Infinitely lived households* or households who treat the well-being of the forthcoming generations in the same way as they treat their own. Real households, however, are mortal, and do not care about future generations as much as they care about themselves or their own children. This is why fiscal policy is effective. This can be formalized in *overlapping-generation models** where individuals make rational decisions over their finite lifetimes whereas the public budgetary constraint holds over an infinite period of time (Blanchard, 1985).

Box 3.6 Ricardian Equivalence

Let us consider an infinitely lived individual, who can freely lend and borrow at a constant interest rate r (cf. Seater 1993). At each period t, this individual receives an income Y_t, consumes C_t and invests (or borrows) $Y_t - C_t$ at the rate r. The intertemporal budget constraint, which states that the present value of all earnings equals the present value of all expenditures, can be written as:

$$\sum_{0}^{\infty} \frac{Y_{t+i}}{(1+r)^i} = \sum_{0}^{\infty} \frac{C_{t+i}}{(1+r)^i} \tag{B3.6.1}$$

Under such constraint, the individual maximizes his/her intertemporal utility function (which we assume as separable, i.e., it is expressed in terms of each period utility):

$$U(t) = \sum_{0}^{\infty} \frac{u(C_{t+i})}{(1+\rho)^i} \tag{B3.6.2}$$

where ρ is the discount rate which measures the individual's preference for the present, and u a concave function. Denoting λ the Lagrange multiplier, the first-order condition of this constrained linear maximization program is:

$$u'(C_{t+i}) = \lambda \left(\frac{1+\rho}{1+r} \right)^i \tag{B3.6.3}$$

Knowing that $u'(.)$ is decreasing, the optimal consumption path of the individual can thus be derived: When $r = \rho$, consumption is kept constant over time; when $r > \rho$, consumption grows over time; when $r < \rho$, it decreases. Contrary to the Keynesian approach, consumption is therefore independent of *current* income. A temporary *fluctuation* of income (a recession, for example) does not affect consumption: The individual will smooth it. However, a permanent *fall* of income over the life cycle (for example, due to a pension reform) reduces consumption, even if current income is unchanged.

Now let us introduce a government, which spends G_t and receives lump sum taxes T_t. We assume that the level of public spending does not enter the individual's utility function (the government does not spend on the provision of public services). First, assume that the government maintains fiscal balance: $G_t = T_t$ at all times. The intertemporal budget constraint of the individual becomes:

$$\sum_{0}^{\infty} \frac{Y_{t+i} - T_{t+i}}{(1+r)^i} = \sum_{0}^{\infty} \frac{C_{t+i}}{(1+r)^i} \qquad \text{(B3.6.4)}$$

The first-order condition of the utility maximization program is unchanged: The introduction of public spending financed by lump-sum taxation does not change the nature of the optimal consumption path of the individual (but the level of consumption falls at each period).

Now let us suppose that the government, while preserving its spending program, decides to reduce taxes at time $t = 0$ by an amount B, and floats securities for this amount B with a maturity of M years at the interest rate r. It is further assumed that interest payments on debt B, as well as the repayment of capital, will be financed by lump-sum taxation. The individual's utility maximization program is unchanged: During the initial period, a part B of income is used to acquire the public debt securities, but taxes also fall by B; during the following periods, until $t = M$, additional income rB is received on securities held, but also additional taxes rB are paid as a result of the need to finance the government debt service; finally, at period $t = M$, the capital B is repaid to the households, but taxes rise by the same amount. The behavior and the level of consumption remain unchanged at each period. The only variable which changes is savings, the difference between disposable income and consumption. At time $t = 0$, the saving of the individual increases by B; the resulting income (at rate r) finances the necessary tax rises, fully expected by the consumer.

This result is the Ricardian equivalence theorem. It expresses the idea that debt is deferred taxation. It also suggests that the central fiscal-policy problem consists in determining the level and nature of public expenditures, more than its financing method (through debt or through taxation).

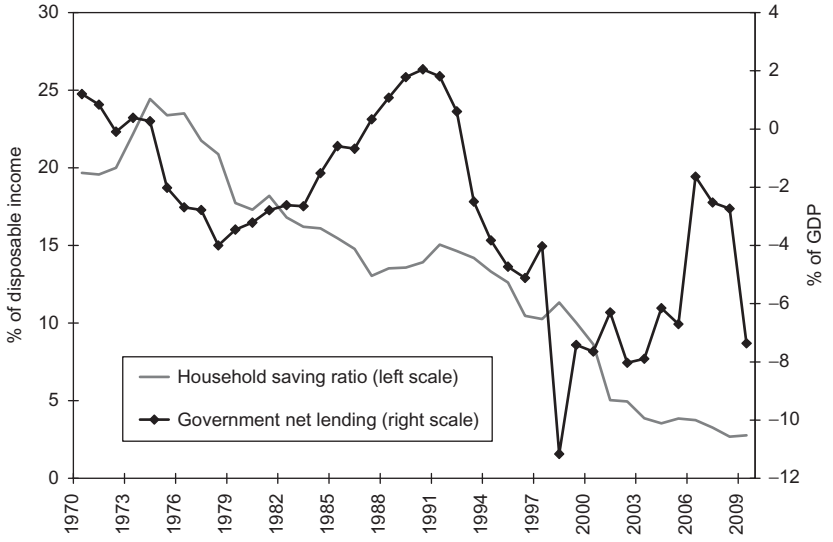


Figure 3.16 Households' and public savings in Japan.
Source: *OECD Economic Outlook* no. 86, November 2009.

A simple and rudimentary check consists in comparing the respective changes in private and public savings over time. Figure 3.16 shows the case of Japan, where a growing public debt presumably made individuals more "'Ricardian." If Ricardian equivalence held, one would observe a perfectly negative correlation between public and private savings, which is not the case. In fact, empirical tests reject full Ricardian equivalence but tend to confirm the reality of some Ricardian effects that reduce the effectiveness of fiscal policy.[21]

c) Empirical assessment of fiscal multipliers

On the whole, the effectiveness of fiscal policy in stabilizing economic activity in the short run is an empirical question that still generates considerable controversy. Because of methodological difficulties, fiscal multipliers are estimated with significant uncertainty. Beyond the typical identification problem that results from distinguishing the effects of discretionary fiscal policy from other factors at play, taking into account dynamic effects is also a challenge, as estimating them depends on assumptions about other policy instruments (namely, interest rates) and economic variables. Because of these difficulties, existing studies unsurprisingly lead to a wide range of estimates

21. For example, de Mello et al. (2004) find, in a sample of 21 OECD countries, that the rise in private saving substitutes for a proportion, ranging from a third to half, the fall in public saving. See also Bayoumi and Sgherri (2006).

for fiscal multipliers, from less than zero to more than four, depending on underlying assumptions.[22]

Empirical studies rely on three different types of models (see box 1.6 in chapter 1):

- Macroeconometric models,
- Dynamic Stochastic General Equilibrium (DSGE) models, and
- Structural Vector Auto Regressive (VAR) estimations.

The first two methodologies generally lead to positive but relatively small (less than unity) multipliers in the short run (one or two quarters), insignificant or even negative ones in the longer term (see, for instance, Hemming et al., 2002, or Briotti, 2005, and the comparison in table 3.4).[23] Structural VAR models often lead to more significant multipliers. For instance Blanchard and Perotti (2002) find significant multipliers for both public spending shocks and net tax shocks in the US. Still, both multipliers are lower than one, even in the short run, and they appear unstable over time.

Following standard Keynesian analysis, a fiscal expansion is more effective when it is carried out through an increase in public consumption or investment rather than through a reduction of taxes or an increase in transfers to households. This is because one euro given to households will not necessarily translate into one euro of additional demand due to the propensity of households to save. Consistently, econometric models generally find lower fiscal multipliers for net taxes than for expenditures, and this is what is found in Blanchard and Perotti (2002). However, allowing for supply-side effects (see below) may change this diagnosis, especially in the most recent periods where households have acquired better opportunities to smooth consumption over time. If a tax cut raises potential output, then this additional permanent income fuels an increase in consumption in the short run that can reinforce the short-run multiplier. Additionally, the rise in potential output can prevent inflation pressures, and hence a crowding-out through the interest rate. Consistently, Mountford and Uhlig (2008) find that tax multipliers exceed spending ones within a structural VAR model.[24]

It remains difficult to provide a general assessment on the impact of fiscal policy since this will depend on the type of tax or spending, on the position of the economy in the business cycle (i.e., whether supply and liquidity constraints are binding or not), on the degree of openness to trade, on the monetary regime, and on the situation of public finances (Ricardian argument). A few lessons emerge, however, from the literature (Spilimbergo et al., 2008, Appendix II): First, government investment

22. A short review is provided in Spilimbergo et al. (2008), Appendix II, pp. 17–21.

23. Cour et al. (1996), however, found large differences across studies, with a significant probability for the multiplier to be *negative*.

24. Bénassy-Quéré and Cimadomo (2006), however, find evidence of large time variations of both spending and net tax multipliers between the early 1970s and the 2000s.

multipliers do not appear to be significantly larger than government consumption multipliers; second, over the medium term, tax multipliers are not necessarily smaller than spending or investment multipliers; third, there is a wide variation across countries, larger countries tending to have larger multipliers. In addition, coordinated fiscal expansion will typically produce larger multipliers. Finally, for a given discretionary fiscal impulse, fiscal multipliers are higher when automatic stabilizers are stronger, that is, in countries where social security is more developed.

3.2.2 Public debt sustainability

Our analysis so far has focused on flows—receipts, expenditure, and deficits. But flows result in stock accumulation, meaning that deficits give rise to debt. Debt, in turn, needs to be serviced, which impacts on deficits. We therefore need to look into the public debt accumulation issue.

a) Solvency

Ricardian equivalence theory emphasizes the government intertemporal budget constraint, which sooner or later calls for raising taxes when spending has increased. Borrowing is only deferring charges to the future. Unlike households, however, governments consider themselves to have an infinite lifetime,[25] so their debt never requires to be redeemed. To be more precise, expiring debt will be paid off through new borrowing, because it is reasonable to think that future generations will be willing, when their turn comes, to use part of their savings to acquire government securities. Is there no limit to the state's borrowing capacity? Asking this question amounts to assessing the state's *solvency** (i.e., the availability of resources allowing it to meet its commitments).

It is relatively easy to determine when a household or a private firm is insolvent but the same does not hold for a government. At first sight, the capacity of a state to ensure the service of its debt could appear unlimited, since it has the power to raise taxes or, if the central bank is not independent, monetize the deficit (which is equivalent to a tax since induced inflation reduces the purchasing power of households). However, even before capacity to pay is exhausted, the political limits of the willingness to pay can be reached.

25. There are examples of states that wind up and close their books, but legacy debt is then carried over to newly established countries. For instance, Czechoslovakian debt was split between the Czech Republic and Slovakia on 1 January 1993. There are also examples of governments that refuse to pay for their predecessors' debt because they deem it politically illegitimate. This famously happened after the Russian and Chinese revolutions. It was actually proposed, as a form of sanction, to formally declare debts incurred by illegitimate dictatorship *odious debt**, meaning that successor governments have a right to repudiate it (Kremer and Jayachandran, 2002). But since political regime change cannot usually be foreseen, it is difficult to integrate it into *ex ante* sustainability analysis.

As illustrated by many historical episodes, from *Ancien Régime* crises to Argentina's bankruptcy in 2002, bankruptcy occurs when citizens no longer accept a further reduction of their income to the profit of the creditors of the state. This is why evaluating the solvency of a state and devising adjustment programs are daring exercises. A senior official for the IMF, John Boorman, expressed it as follows:

> Debt can almost always be serviced in some abstract sense, through additional taxation and through the diversion of yet more domestic production to exports to generate the revenue and foreign exchange needed to service the debt. But there is a political and social, and perhaps moral, threshold beyond which policies to force these results become unacceptable.
>
> J. Boorman (2002)

Another difference between a state and a private borrower is that there is no collateral for sovereign debt. If a state defaults on its commitments, neither domestic nor foreign creditors can seize its assets (unless the latter invade the country).[26] An indebted state's attitude toward its creditors depends on the benefits and costs of defaulting on its debt. The benefits result from writing off the debt and the corresponding interest burden, while the costs are mainly reputational: A defaulting state may be cut off from financial markets or at least pay a higher risk premium in the future. History however shows that defaults are frequent, and that especially in recent times, states rather quickly regain access to financial markets (Reinhart and Rogoff, 2008). Unlike for private creditors, assessing the solvency of a state requires an evaluation of its willingness to pay.

If resources exist but cannot be mobilized immediately (one can think of forthcoming fiscal receipts or of state-owned companies that cannot be sold immediately due to lack of purchasers), or if they are available but can dry up at short notice (such as short-term credit lines extended by foreign banks), there is a risk that the government defaults even though it is solvent: This is a *liquidity crisis**.[27]

b) From solvency to sustainability

Solvency characterizes the situation of public finance at a given moment in time, but in view of the inertia of public expenditures and receipts (in no large country can spending be cut by 10% of GDP from one year to another, for example), it is always important to be able to anticipate possible insolvency at any future time. This is what the concept of *sustainability** addresses.

Public finance is said to be *unsustainable** if, on the basis of the current economic policy and of available forecasts, the expected development of the public debt leads inevitably to a situation of insolvency. Fiscal policy can

26. To be more precise, they can seize some of its foreign assets but those generally amount to a small fraction of liabilities.

27. See section 2 of chapter 4 for a theoretical discussion of liquidity in the case of banks.

therefore be unsustainable without solvency problems arising immediately. Nonetheless, this policy will have to be modified in the future. Examples abound: In 2003–04, no one would question the solvency of the US Federal Government, but the fiscal policy of the Bush administration was considered unsustainable by numerous observers (Auerbach et al., 2004). In August 2003, the International Monetary Fund thus concluded its annual review of the United States:

> [IMF directors] stressed, however, that for the economy's full potential to be realized, decisive action will need to be taken over the coming years to re-establish a strong US fiscal position. In particular, they expressed concern that the worsening of the longer-term fiscal position, including as a result of the recent tax cuts, will make it even more difficult to cope with the aging of the baby-boom generation, and will eventually crowd out investment and erode US productivity growth.
>
> International Monetary Fund (2003)

Public finance sustainability is especially important in a monetary union where the central bank is independent, as is the case in the euro area. Suppose that a member state cannot service debt (interest and principal). Since it cannot rely on monetization by the central bank, there are three options: (i) A massive adjustment combining cuts to primary expenditures and tax increases; (ii) temporary support by other member states and the International Monetary Fund; or (iii) a partial default whereby the government negotiates a debt reduction with its creditors. The second option addresses the short-run solvency problem, but not the sustainability one, since emergency support is by nature temporary and will need to be refunded.[28] As for the latter one, by devaluing banks' assets (which include many government securities), it is likely in turn to cause banking crises. The final outcome could be a takeover of ailing banks by foreign banks or, in the worst case, an *ex post* monetization by the central bank.

More generally, solvency crises are rather frequent events, as documented by Reinhart and Rogoff (2009). In practice, they are generally solved by a combination of the three options. Governments request the assistance from international financial institutions, essentially the IMF. When assisting a country, the IMF however requires the government to devise and implement an *adjustment program** aimed at restoring external debt sustainability.[29]

28. The EU Treaty (Art. 125 of the Treaty on the functioning of the EU) prohibits member states or the EU as a whole taking responsibility for a member country's debt. In 2010, the members of the euro area nevertheless decided to extend medium-term facilities to Greece. Whether this decision was consistent with Art. 125 was intensively debated, especially in Germany. The EU's line of defense was that this would not make it liable for the commitments of the Greek government, which is consistent with the no bail-out clause.

29. Not all crises are triggered by concerns about solvency. Some are pure liquidity crises, e.g. sudden stops in market financing. This is why the IMF introduced in 2009 an insurance facility, the *flexible credit line**, available to countries which are solvent but vulnerable to liquidity crises.

Assistance, adjustment and rescheduling are often not sufficient to restore sustainability, which leads the government to negotiate a debt reduction with its public and private creditors. These negotiations take place in the Paris club (for official creditors), the London club (for banks) and ad-hoc fora (for nonbank private creditors).

c) Assessing debt sustainability

There is no universal criterion for assessing public debt sustainability. A first, very rough one relies on the stability of the debt-to-GDP ratio. Consistently, the observed primary balance is compared to the primary balance that would allow the debt ratio to stay constant, called the *debt-stabilizing deficit**. The latter depends on the debt ratio and on the difference between GDP growth and the interest rate, as shown in box 3.7.

A simple application based on 2009 data is provided in the box for a few advanced countries. The global crisis brought GDP growth rates lower than interest rates, requiring primary surpluses to stabilize debt ratios. However, governments ran primary deficits as an attempt to stabilize their economies.

The problem with this first approach is that the observed debt-to-GDP ratio may not correspond to an optimal, long-run level. The debt ratio of the Czech Republic jumped from 12.2% to 37.6% of GDP between 1997 and 2003: Was it problematic for a moderately indebted country engaged in a full transition toward a market economy to increase its public debt ratio in order to finance investments in infrastructures and structural reforms conducive to growth? The observed level of debt is not necessarily optimal, so stabilizing the debt ratio at its current level may provide inadequate guidance to debt sustainability concerns.

Box 3.7 How to Stabilize the debt-to-GDP ratio

Here we start from box 3.3 that describes debt accumulation as the following process:

$$b = \frac{(1+i)}{(1+n)} b_{-1} + d \cong (1 + i - n) b_{-1} + d$$

$$\cong (1 + r - g) b_{-1} + d \qquad \text{(B3.7.1)}$$

Again, we neglect market valuation and all stock adjustments such as privatizations. A rough approach to sustainability then requires the ratio of public debt to GDP to be constant: $b = b_{-1}$. To obtain this stability, the primary deficit needs to be:

$$d = \frac{n - i}{1 + n} b \cong (n - i) b \cong (g - r) b \qquad \text{(B3.7.2)}$$

And the financial deficit:

$$d + ib \cong nb. \qquad \text{(B3.7.3)}$$

For a debt ratio of 60% of GDP and a nominal growth rate of 5% (namely 3% of real growth plus 2% inflation), the financial deficit consistent with a constant debt ratio is 3% of GDP. This is where the fiscal discipline criteria imposed in the Maastricht Treaty come from. Moreover, for a real interest rate of 2%, the primary deficit compatible with the stability of the debt ratio at 60% of GDP amounts to 0.6% of GDP. Conversely, the primary balance has to be in surplus when the real interest rate exceeds the real growth rate. Such a situation prevailed in Europe in the 1980s and 1990s. Countries such as Italy and Belgium had to run considerable primary surpluses (negative primary deficits) in order to reduce their public debt ratios. Table B3.7.1 provides an illustration of debt-sustainability assessment along this simple arithmetic in the wake of the 2007–09 crisis. For instance, stabilizing the Greek debt ratio at its end-2009 level would have required a 3.6%-of-GDP primary surplus in 2010–11, whereas the OECD was forecasting a 5.2%-of-GDP primary deficit at that time.

Table B3.7.1
Stabilizing the debt-to-GDP ratio: Short-term exercise from 2009

	Gross debt b (% of GDP) End 2009	Nominal growth n (% per year) Avg 2010–11	Long-run nominal interest rate i (% per year) Avg 2010–11	Required primary deficit[a] $b(n-i)$ (% of GDP) Avg 2010–11	Observed primary deficit[a] d (% of GDP) Avg 2010–11
Austria	70.3	3.0	4.0	−0.7	2.3
Belgium	101.0	3.2	4.0	−0.9	1.1
Denmark	51.8	3.6	4.1	−0.3	4.2
France	86.3	2.8	4.1	−1.1	5.0
Germany	76.2	2.4	3.8	−1.1	2.6
Greece	119.0	−2.6	7.1	−11.5	2.0
Ireland	70.3	0.0	5.3	−3.7	8.3
Italy	128.8	2.2	4.6	−3.1	0.4
Japan	192.9	1.2	1.9	−1.4	6.5
The Netherlands	68.6	2.5	4.0	−1.1	4.0
Portugal	87.0	1.9	4.9	−2.6	2.9
Spain	62.6	0.5	4.4	−2.4	6.7
Sweden	31.8	5.0	3.8	0.6	0.9
UK	72.3	3.7	4.7	−0.7	8.8
US	83.0	4.2	4.7	−0.4	7.7
Euro area	86.3	2.2	4.2	−1.7	3.4

[a] A positive figure points to a primary deficit.

Source: OECD, Economic Outlook No. 87 forecasts (April 2010) and authors' own calculations.

Sustainability is difficult to define as it should take into account the possibility of a state remaining permanently in debt (because it is infinitely lived) but must exclude "pushing the debt ahead" as in speculative chains or *Ponzi games**.[30] The technique was made famous again by Bernard Madoff's misdeeds uncovered in 2008, but it has been known for a long time: Lewis Carroll pleasantly illustrates it in *Sylvie and Bruno*:

> "Ah, well, I can soon settle his business," the Professor said to the children, "if you'll just wait a minute. How much is it, this year, my man?" The tailor had come in while he was speaking.
>
> "Well, it's been a doubling so many years, you see," the tailor replied, a little gruffly, "and I think I'd like the money now. It's two thousand pound, it is!"
>
> "Oh, that's nothing!" the Professor carelessly remarked, feeling in his pocket, as if he always carried at least that amount about with him. "But wouldn't you like to wait just another year, and make it four thousand? Just think how rich you'd be! Why, you might be a King, if you liked!"
>
> "I don't know as I'd care about being a King," the man said thoughtfully. "But it does sound a powerful sight o' money! Well, I think I'll wait—"
>
> "Of course you will!" said the Professor. "There's good sense in you, I see. Good-day to you, my man!"
>
> "Will you ever have to pay him that four thousand pounds?" Sylvie asked as the door closed on the departing creditor.
>
> "Never, my child!" the Professor replied emphatically. "He'll go on doubling it, till he dies. You see it's always worthwhile waiting another year, to get twice as much money!"
>
> Carroll (1889), quoted by Keynes (1931)

A second, more rigorous definition of sustainability starts from the government's intertemporal budget constraint: Public finance is deemed sustainable if the present value of all future public receipts is at least equal to the present value of future spending plus the initial value of outstanding debt (cf. box 3.8).

Consistently, the sustainability of public finance can be assessed by comparing the global tax pressure with the *sustainable tax rate** that ensures debt sustainability, for a given path of public expenditures and depending on assumptions about growth and interest rates. This approach is now used in the EU to monitor the fiscal position of member countries in the framework of the Stability and Growth Pact (see box 3.14), as a complement to debt and deficit analysis. Based on long-run projections on public expenditures (especially those related to health and aging), the European Commission (2009) has proposed a numerical application. The results are reported in the second

30. From the name of a famous Boston crook in the 1920s, who used to entice savers with the promise of high returns, but who would pay them only with the amounts collected from new participants. Ponzi games were played on a large scale in Russia and Albania in the 1990s. The Madoff fraud uncovered in 2008 is another example. All these games, however, always end in similar ways.

Table 3.5
Increase in tax pressure necessary to fulfill alternative sustainability criteria (% of GDP)

	Debt/GDP = 60%	Intertemporal budget constraint (infinite horizon)
Germany	3.1	4.2
France	5.5	5.6
UK	10.8	12.4
Italy	1.9	1.4
Greece	10.8	14.1
Spain	9.5	11.8
Poland	2.9	3.2
Portugal	4.7	5.5
Hungary	−1.1	−0.1
Euro area	4.8	5.8

Source: European Commission (2009).

column of table 3.5. In 2009, the increase in the tax pressure required to meet the intertemporal budget constraint was for most countries more demanding than that required to return to the debt threshold of 60% set by the EU Treaty.

This approach, of course, is fragile in that it relies on long-term projections of growth, interest rates, and especially public expenditures. Furthermore, it provides a global assessment of debt sustainability but does not give any clue as to what the adjustment path should be. Finally, it should be noted that the sustainable tax rate can "jump" in response to a change of economic policy scenario—for example, a pension reform which reduces future government spending relaxes instantly the sustainability constraint.

Box 3.8 The Mathematics of Debt Sustainability

Since states do not have a predefined, finite lifetime, there is no need for the net public debt to fall to zero at a given date in the future. Rather, debt sustainability implies that *the present value of debt at time t tends toward zero as t tends to infinity*. This condition, called the *transversality condition**, is equivalent to the equality between the present value over time of the government future income and expenditure streams corrected for the initial level of debt.[31] Note that it does not imply that the debt ratio

31. The layman's version of the transversality condition is Herbert Stein's famous remark that "if something cannot go on forever, it will stop."

goes to zero when time t tends to infinity, since a nonexplosive debt ratio is consistent with sustainability.

We start with the continuous-time equivalent to the debt accumulation equation of box 3.3:

$$\frac{db}{dt} = (i - n)b + d = (r - g)b + d \qquad \text{(B3.8.1)}$$

The variation of the debt ratio b is a function of the interest rate (i in nominal terms, r in real terms), of the growth rate (n in nominal terms, g in real terms) and of the primary deficit d. Assume, for the sake of simplicity, that the real interest rate and the growth rate are constant, and let b_0 represent the initial debt ratio. The debt ratio at time t can be obtained by integrating (B3.8.1) over time:

$$b_t = b_0 e^{(r-g)t} + \int_0^t d_s e^{(r-g)(t-s)} ds \qquad \text{(B3.8.2)}$$

The present value of b_t at $t = 0$ is obtained by multiplying both sides of this equation by $e^{-(r-g)t}$. The discount rate $(r - g)$ allows taking into account the dampening effect of growth on the debt ratio.

$$b_t e^{-(r-g)t} = b_0 + \int_0^t d_s e^{-(r-g)s} ds \qquad \text{(B3.8.3)}$$

When t tends to infinity, the present value of the debt ratio has to tend toward zero, which implies that the right-hand-side of the equation also tends toward zero:

$$\lim_{t\to\infty} b_t e^{-(r-g)t} = 0 \quad \text{implying} \quad b_0 = -\int_0^\infty d_s e^{-(r-g)s} ds \qquad \text{(B3.8.4)}$$

The first condition is called the transversality condition. If $r > g$, it is necessary and sufficient that the debt ratio increases at a lesser pace than the discount rate $r - g$. If $r < g$, the government can finance the debt service through new borrowing while remaining solvent. This was the situation of the 1970s, a period when public debt problems were benign. But in the 1980s and 1990s, in Europe, real interest rates were higher than the growth rates of the economy.

The second condition implies that the present value of future primary surpluses (the opposite of the primary deficits) "repays" the initial debt. Writing $d = x + h - \tau$, where x designates expenditure on goods and services, h public transfers, and τ taxes and levies, the condition becomes:

$$b_0 + \int_0^\infty (x + h)_s \, e^{-(r-g)s} \, ds = \int_0^\infty \tau_s \, e^{-(r-g)s} ds \qquad \text{(B3.8.5)}$$

The sum of the initial debt and of the present value of future expenditures has to equal the present value of future income streams. This is the *intertemporal budget constraint of the government**.

Blanchard (1993) takes the sequence of public expenditures and transfers as given in terms of GDP, and calculates the constant tax rate τ^*, which he calls the *sustainable tax rate*, that ensures debt sustainability:

$$\tau^* = (r - g)\left[b_0 + \int_0^{\infty} (x_s + h_s)e^{-(r-g)s} ds\right] \quad \text{(B3.8.6)}$$

τ^* is therefore the rate of taxation sufficient to service (at rate $r - g$) the sum of the initial debt and of the present value of the prospective stream of expenditures on goods and services and on transfers. The gap between the sustainable tax rate τ^* and the observed tax rate τ provides an indicator of sustainability. If $\tau < \tau^*$, the long-term sustainability of public debt requires either a rise in the tax rate τ, or a cut in expenditures on goods and services x or on transfers h.

To allow the calculation of measurable indicators on the basis of this theoretical approach, Blanchard proposes calculating the constant tax rate necessary to restore the initial level of the ratio of public debt after a given number N of years:

$$\tau_N^* = (r - g)\left[b_0 + \frac{1}{1 - e^{-(r-g)N}} \int_0^{N} (x_s + h_s)e^{-(r-g)s}\, ds\right] \quad \text{(B3.8.7)}$$

The sustainable tax rate is still the rate that makes it possible to cover the present value of the foreseeable expenditure and the interests on the initial debt, but the expenditure stream taken into account now refers to the period under consideration only, i.e., from 0 to the year N.

A third approach to debt sustainability is backward-looking. It consists in analyzing, on the basis of past observations, the joint dynamics of the deficit and debt to evaluate the likelihood of diverging scenarios that would violate the "transversality condition" of box 3.8. This method amounts practically to testing the existence of a systematic force pulling the tax pressure back toward the expenditure-to-GDP ratio (cf. box 3.9). This approach offers an assessment of sustainability independently from any long-run forecast. However, because it is backward-looking, it cannot evaluate the impact of a reform, such as a pension reform, on debt sustainability.

Box 3.9 An Econometric Approach to Debt Sustainability: The Hamilton and Flavin Method

A shortcoming of the previous approaches is that they ignore uncertainty. However, debt can increase under the effect of economic shocks (e.g., recessions), fiscal shocks (e.g., falls in the tax yield), or wealth shocks (e.g., asset depreciation, as, for example, when public sector companies incur losses). The experience of sovereign-debt crises in emerging countries has shown the extent to which taking into account such shocks can result in a different assessment of sustainability. Hamilton and Flavin (1986) propose an alternative method for assessing sustainability under uncertainty. They rely on the debt-accumulation equation of box 3.3, assuming constant interest rate and constant growth rate in the long run. The expected variation of the debt ratio is the following:

$$E_t b_{t+1} - b_t = E_t d_{t+1} + (r-g) b_t \quad \text{hence} \quad b_t = -\beta E_t d_{t+1} + \beta E_t b_{t+1} \tag{B3.9.1}$$

Where $\beta = \frac{1}{1+r-g}$ and $E_t(X)$ denotes the expected value of X, conditional on information available at date t.

Now let us consider $\varepsilon_t = b_t - \beta b_{t+1} + \beta d_{t+1}$. We have $E_t \varepsilon_t = 0$ but due to shocks to the primary deficit d_t, ε_t is uncertain. This leads to an empirical definition of sustainability: The debt is said to be sustainable if ε_t is stationary, i.e., of constant mean and of variance limited over time, an hypothesis that can be tested.

Boissinot et al. (2004) applied a similar method to analyzing the French situation, by using the following equation:

$$\tau_t = \alpha + \beta(x_t + h_t + i_t b_{t-1} + u_t) \tag{B3.9.2}$$

Where u_t is the error term and other variables are the same as in box 3.8. If there is a long term relation with $\beta = 1$, then a permanent increase in spending induces an identical increase in tax receipts, the public deficit is stationary, and the present value of debt tends toward zero as t tends to infinity. This situation is described as "strong sustainability." If there is such a relation with $0 < \beta < 1$, tax receipts increase less quickly than spending and the debt ratio increases over time. However, the transversality condition still holds if β is strictly positive, because the increase in spending eventually results in an increase in tax receipts. This situation is described as "weak sustainability." Boissinot et al. (2004) found a coefficient β of 0.24 over the period 1978–2003, corresponding to weak sustainability. This coefficient has substantially deteriorated since the early 1990s, when it was equal to about 0.5.

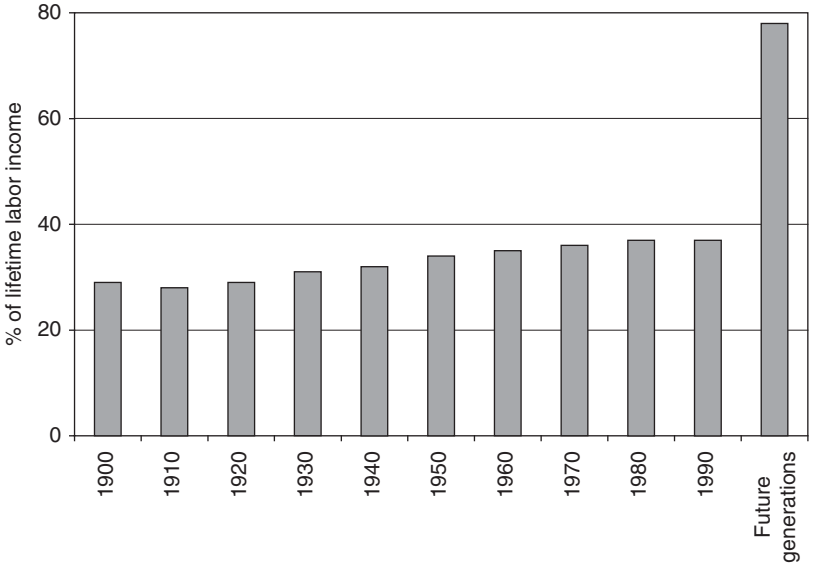


Figure 3.17 Generational accounting: Estimated lifetime net taxes by year of birth (averages in %).
Source: "Who Pays and When. An Assessment of Generational Accounting," Congressional Budget Office, 1995, based on data from Auerbach et al., 1991.

A fourth approach focuses on the intergenerational dimensions of fiscal policy. The so-called *generational accounting** approach compares the present value (at the time of birth) of taxes net of government transfers for present generations as well as for the newborn. It provides a useful analytical tool for assessing who pays for what and when, and who transfers what to whom, and it is also relevant for assessing the sustainability of a given policy (Auerbach et al., 1991). The sustainability criterion then stipulates that the present value of net taxes paid by future generations should be equal to the sum of the current debt and of the present value of all government spending less the present value of net taxes paid by the present generation.

Figure 3.17 provides an example of such reasoning for the US. The graph shows a gradual rise in each generation's net contributions as a percentage of their labor income. More importantly, it shows that the present legislation leaves to future generations a burden that is about twice the contribution of living generations.

On the whole, there still remains a gap between the theoretical and empirical approaches. The latter suffer from the absence of data of sufficient quality on public accounts, and of a dependency on the models and the assumptions used. The indicators informing economic policy-making therefore remain very rudimentary. This provides no reason for not organizing

the discussion precisely and for not monitoring the consistency between spending and income projections. International discussions on fiscal policy are increasingly focusing on debt sustainability: Examples include the revised Stability and Growth Pact in the euro area, further discussed below, and the "debt sustainability framework" developed by the IMF and the World Bank to gauge the capacity of low-income countries to take new loans and repay them without getting trapped in a new debt crisis.

d) The political economy of debt

One specificity of fiscal policy is that it may provide benefits in the short run while reducing the room for maneuver of future governments, or even future generations, who will have to face an inflated public debt. This intertemporal feature has implications for policymaking. It is the task of political economy to uncover them.

Box 3.10 provides an example of a model where the level of public debt is chosen by voters based on the distribution of wealth across voters, since wealthy voters are those who hold government bonds: Wealthy voters and their children (who receive bequests) are in favor of public debt to the extent that it is repaid to them; poor voters prefer not to pay taxes to repay the debt, hence they prefer either no debt or a repudiation of the debt. The government will run into debt, and will repay it if there is a constituency of (relatively wealthy) bondholders.

Other political-economy or credibility-based approaches study, for instance, how a partisan majority can constrain its successors by financing its priorities (or by preventing them from financing theirs), or how the structure of the public debt (its maturity, its currency composition, or whether it is indexed to inflation or not) signals the government's intentions as regards economic policy.[32] For instance, issuing inflation-indexed debt can strengthen the government's commitment to fight inflation since interest costs on existing debt will go up with inflation.

Box 3.10 The Politics of Debt According to Tabellini (1991)

Guido Tabellini asks why government debt, which shifts the tax burden to future generations of taxpayers, is eventually repaid even though it is created without the consent of those who will bear the burden. His model is an overlapping-generations, two-period model of a closed economy. Only one generation, the "parents," is present in period 1. In period 2, a new generation of children is born. Parents live two periods, children

32. See Elmendorf and Mankiw (1999) for a literature review.

only one. Families are connected by altruistic links: Parents care about the well-being of their children (and can leave them bequests), and reciprocally. In period 1, wealth is unequally distributed. However, the children's income in period 2 is uniform, which echoes the observed fact that wealth inequalities are larger than income inequalities. Parents vote (under majority rule) in period 1 on how much debt to issue. The key of the model is the possibility, in period 2, of defaulting on part of this debt: The proportion of debt subject to default is determined in period 2 by a vote (under majority rule) in which parents and their children take part. The remainder of the debt is refunded by a tax on children.

The incentives faced by each individual become clear when one realizes that repudiating the debt redistributes wealth from the rich toward the poor, as a progressive tax system would do. In period 2, only the rich and their children have an interest in having the debt repaid. Highly unequal wealth distribution therefore leads to a high default rate on debt. In period 1, incentives are more complex: Parents tend to profit from the fact that their children do not vote to float a large quantity of debt and thus present them with a *fait accompli*—but lenders have to take into account the possibility that a fraction of this debt will be repudiated. If the poor are numerous (wealth is very concentrated), one knows in advance that a major part of the debt will be repudiated in period 2, so that nobody is ready to lend in period 1. In contrast, if wealth is evenly distributed, a larger quantity of debt can be issued.

Both decisions about debt emission and debt repayment reflect the structure of incentives in both periods. On the whole, the relation between wealth inequality and the size of the debt is not monotonous. If inequality is high and the rich are a minority, the latter will be spoiled and debt is politically impossible. However, if equality prevails, no child is willing to repay the debt, and debt is also impossible. Summing up, Tabellini shows that redistribution through debt between generations is politically viable only when wealth inequalities are neither too weak nor too strong.

Similar models were used to study the repudiation of debt in emerging countries (Bulow and Rogoff, 1989).

3.2.3 Supply-side effects and reconciliation attempts

So far, we have explained how fiscal policy can be expected to affect output in the short run, and we have enumerated several factors—propensity to save or to import, interest-rate or exchange-rate crowding-out, rational expectations—that could reduce the short-run impact of fiscal policies. Next, we have explored the concept of debt sustainability and suggested how public debt can be used strategically. All these clouds that accumulated over the efficient use of fiscal policy led to some discredit of this type of counter-cyclical

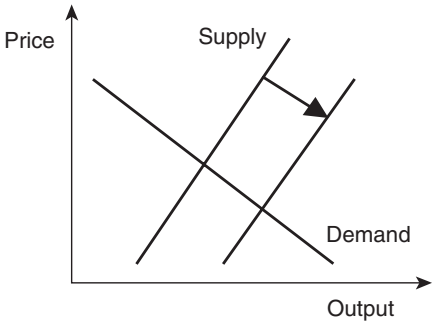


Figure 3.18 Supply-side effects of a tax cut.

policy in the 1980s and 1990s. This was a period when fiscal policies across the world should have been devoted to ensuring debt sustainability. Obviously, this was not the case (remember figure 3.10). This is because tax cuts were then believed to have a positive, long-run impact on growth through supply-side effects.

a) Keynes under attack

For the reasons listed in section 3.2.1, neoclassical (and "new-classical") economists generally deny any significant impact of counter-cyclical fiscal policies. However, they underline the usefulness of a tax cut to stimulate aggregate supply and hence raise potential output: In the AD–AS representation, a tax cut moves the supply curve downward (it reduces the output price for any production level), which stimulates the activity and causes prices to decline, as shown in figure 3.18. Thus, neoclassical economists join the Keynesians in recommending tax cuts when growth is mediocre; but the neoclassical view is that these stimulate supply, while for the Keynesians, they boost domestic demand through the induced rise in disposable income.

As for public spending, the disagreement between neoclassical and Keynesian economists is maximum. The former deny any positive short-run effect of public spending while emphasizing its implications in terms of future rises in taxes which, if rationally anticipated, have a negative short-term impact on consumption. Conversely, they applaud spending cuts because they pave the way for tax cuts that are favorable to long-term growth and in turn force further spending cuts:

> "We didn't starve the beast," laments a White House official. "It's still eating quite well—by feeding off future generations."
>
> Paul Blustein, "Reagan's Record," *The Wall Street Journal*, 21 October 1985

However, neoclassical economists agree with Keynesians not to balance the budget at every point in time, but rather to let the public balance go into

deficit in a recession (and into surplus in a boom). As observed by Robert Barro (1979), because taxes are distortionary, it is not optimal to raise tax rates when tax receipts are affected by a recession, and it is preferable to keep them constant over the cycle. *Tax smoothing**, as it is known, thus results in a prescription similar to that of the Keynesian advocacy of letting automatic stabilizers play in full, but on very different grounds.

b) Non-Keynesian effects

A number of models were proposed in the 1990s to go beyond standard controversies and try to reconcile the apparently contradictory facts mentioned in section 3.1. Rather than building a general model of fiscal policy effects, they aimed at providing a framework in which Keynesian, non-Keynesian (when fiscal expansion has no effect), and anti-Keynesian (when the multiplier becomes negative) behavior could be explained. Starting from different premises, these models suggested that the economy could be Keynesian in normal times, but non-Keynesian or anti-Keynesian in specific budgetary circumstances. In particular, large-scale fiscal adjustments would more likely result in non-Keynesian behavior, because they generally take place during critical periods when agents' expectations are changed.[33]

A first series of models (*neoclassical models with composition effects*) builds on the neoclassical framework, but brings two additional features (Blanchard et al., 1991; Alesina and Perotti, 1995; Perotti, 1996). The first one introduces fiscal distortions, implying that a tax rise (or a spending rise, since a permanent increase in expenditures generates expectations of future tax rises) reduces output through supply-side effects. Under this assumption, the key variable is the permanent public expenditure level. Large-scale fiscal policy changes, which are likely to have a permanent effect on the expenditure level, can therefore have an impact on output. The next step, and it is the second addition, is to assume that in normal times fiscal adjustments generally take the form of tax increases (which validate a pre-established expenditure level, but do not affect it), while periods of fiscal distress more often lead to permanent spending cuts, and are therefore likely to have positive effects on supply.

However, these models with composition effects (between income and spending) are rather extreme in that they can produce non-Keynesian or anti-Keynesian effects, but never Keynesian effects that can nonetheless still be observed in reality.

The second category of models (*Keynesian models with threshold effects*) also rests on the introduction of nonlinearities, but they are built on Keynesian assumptions. The accumulation of public debt was suggested by Blanchard et al. (1991) as the key mechanism. As long as agents believe that public debt remains sustainable, they can ignore its consequences, find it

33. See Giavazzi et al. (2005) for empirical evidence of such nonlinearities.

acceptable that they will be borne by future generations, and adopt a non-Ricardian behavior. But if the debt reaches some critical level, and if its monetization or its repudiation are ruled out, they know that a stabilization program must happen shortly. In the event of an expected tax rise, they save accordingly; in the event of permanent fall in expenditure, which will improve their intertemporal wealth, they start to consume (cf. Bertola and Drazen, 1993). For some debt levels, a negative (anti-Keynesian) correlation will be observed between public and private savings. At some other debt levels, a positive ("pseudo-Keynesian") correlation will obtain.

Sutherland (1997) introduces uncertainty regarding the intergenerational distribution of future taxes. In his overlapping-generations model, presented in box 3.11, consumers have a finite horizon; agents behave in a Keynesian way as long as the public debt remains rather weak so that the burden of fiscal adjustment can safely be transmitted to future generations; they become increasingly anti-Keynesian as the likelihood increases that they themselves will have to support the corresponding burden. The same fiscal impulsion can now lead to opposite results. Such models seem especially relevant to describe situations of fiscal crisis, during which expectations take a prominent role.

Such ideas also find their way into economic policy discussions and statements, as illustrated by these remarks by Jean-Claude Trichet in 2003 when he was appointed President of the European Central Bank:

> . . . there is, in any economy, a threshold. When you cross the threshold, the potential positive Keynesian effects of additional public spending and deficits are offset by what I would call Ricardian effects—namely that you are losing more as regards the confidence of households and of entrepreneurs than you could gain with Keynesian effects. That is why there are always limits to what one can do, the limit has to be judged. It has been judged in Europe in terms of this threshold of 3%. . . .
>
> Jean-Claude Trichet, hearing of 11 September 2003 by the Committee on Economic and Monetary Affairs of the European Parliament (http://www.europarl.europa.eu/hearings/20030911/econ/cre.pdf, p. 4)

Various contributions have sought to test a number of hypotheses likely to explain the anti-Keynesian character of certain large-scale fiscal adjustments, such as the size and the external openness of the country, the policy mix, the credibility gains from a restoration of sound public finance, or households' savings behavior. But they have not led to general results.

Box 3.11 The Impact of Public Debt on Fiscal Policy Effectiveness

Here we present a model due to Sutherland (1997), in which fiscal expansion exhibits the traditional Keynesian effects at moderate public debt levels, because consumers consider that the implied tax burden will

be borne by later generations. Conversely, when debt reaches very high levels, a fiscal expansion may well lead to a *contraction* of output, because consumers anticipate that adjustment will have to take place in their lifetime and expect an offsetting tax increase in the immediate future.

The representation of households' behavior is based on an overlapping-generation model. At any date, two generations coexist, the "young" and the "old." Fiscal policy is represented by a primary deficit D (per capita) that takes the form of a lump-sum transfer toward the consumers (as usual, the lump-sum character of the transfer allows fiscal distortions to be ignored). Denoting r the constant interest rate, the dynamics of the per capita public debt B at time t is given by:

$$dB_t = rB_t dt + D_t \tag{B3.11.1}$$

where the measure of the deficit D_t includes a stochastic component. Under these conditions, the debt could become explosive. To respect the intertemporal budget constraint, Sutherland imagines a discrete adjustment process: When debt reaches a per capita ceiling U, a lump-sum tax of a per capita level T is levied, which reduces the debt to $U - T$; when it reaches a floor (intuitively negative) L, a per capita lump-sum transfer T is paid to the inhabitants, which increases the public debt to $L + T$. Consumers have a finite life, with a constant death probability θ. Each individual consumes a quantity c_t of the same homogeneous good, freely exchanged at a fixed price and derives an instantaneous utility $u(c_t)$, where u is quadratic. The individual receives a fixed income y, plus an income from his/her wealth A, which is placed with insurance companies that inherit in case of the individual's death. Hence, the return on the individual's assets is $r + \theta$. The risk premium θ can be interpreted as a transfer from the consumers who die to the consumers who survive. The consumer's budget constraint is therefore described by:

$$dA_t = [y_t - c_t + (r + \theta)A_t]dt + D_t \tag{B3.11.2}$$

Under this constraint, the following expected utility is maximized:

$$E_t \int_t^\infty u[c_\tau] e^{-(r+\theta)(\tau - t)} d\tau \tag{B3.11.3}$$

Consumption can then be derived as:

$$c_t = y_t + (r + \theta) \left[A_t - E_t \int_t^\infty \delta_\tau T e^{-(r+\theta)(\tau - t)} d\tau \right] \tag{B3.11.4}$$

where the function δ_t takes the value $+1$ when a crisis bringing a debt reduction occurs, -1 when in contrast debt reaches the floor L, and 0 in other cases. In other words, the consumer consumes the flow of income plus the interest income received on wealth, net of the present value

(discounted at the rate $r+\theta$, to take into account the finite life) of expected future taxes.

The results of the model depend on the dynamics of the term:

$$S_t = E_t \int_t^{\infty} \delta_\tau T e^{-(r+\theta)(\tau-t)} d\tau \qquad \text{(B3.11.5)}$$

Sutherland shows that S is an increasing function in B, that $\partial S/\partial B$ is close to zero when B is low (in absolute value), but that $\partial S/\partial B$ is greater than unity when B (in absolute value) approaches the thresholds L or U. When B is low, a fiscal expansion (positive D) increases the consumption of each individual and total consumption. It therefore exhibits the traditional Keynesian effect. When B approaches U, the same deficit D generates expectations of an impending adjustment and causes a *reduction* of individual and total consumption in preparation for the tax increase to come: In that case, a fiscal expansion therefore exhibits an anti-Keynesian effect and leads to a contraction of output.

Table 3.6 summarizes the expected effects of a fiscal contraction according to the various theoretical frameworks.

3.3 Policies

As described in the previous section, fiscal policy faces considerable uncertainty. After a period of widespread conviction about the quasi-mechanical effects to be expected from fiscal policy, a more moderate approach has settled in, that qualifies the Keynesian vision of the fiscal multiplier. The increasing relevance of debt sustainability issues and the awareness of the role of private agents' expectations in the transmission of fiscal policy have led to substantial refinements of the analysis. Credibility and reputation problems have surfaced in fiscal policy, as in other economic policy areas. In particular, governments' commitments suffer from a *time inconsistency problem**.[34] To generate expectations favorable to private demand, the government may announce a virtuous policy consisting, for example, in maintaining fiscal balance; but over time it faces incentives to renege on its commitment in order to lift output. Recognizing these incentives, private agents have no

34. The "time inconsistency" problem refers to the fact that multi-year commitments announced by a government in order to maximize a social utility function over time do not necessarily correspond to the choice of policies that would emerge from a repeated maximization allowing a government to determine the optimal policy period after period—see chapter 2 for a general presentation and chapter 4 for a discussion in the case of monetary policy.

Table 3.6
Effect of a restrictive fiscal policy within various theoretical frameworks

	Hypotheses	Mechanisms	Effect of a fiscal contraction
Neo-Keynesian models	Short–medium-term horizon. Flexible supply conditions.	Partial financial crowding-out. Absence of nonlinearities. ***KEYNESIAN***	Recessionary
Ricardian equivalence	Intertemporal budget constraint. Consumers with infinite horizon. Rational expectations.	Crowding-out one for one of private consumption by public consumption. Neutrality of the deficit. ***NON-KEYNESIAN***	Neutral
Neoclassical models with composition effects	Neo-Ricardian framework. Fiscal distortions. The composition of the adjustment depends on the initial conditions (debt level . . .)	Super-crowding-out due to supply-side effects. ***ANTI-KEYNESIAN***	Expansionary (if poor initial conditions, i.e., high debt)
Keynesian models with threshold effects	Keynesian rigidities. Consumers with finite horizon. Probability of "stabilization" grows with the debt.	Keynesian mechanism under standard conditions. Inversion of the effects under poor public finance situation. ***KEYNESIAN*** or ***ANTI-KEYNESIAN***	Recessionary if debt is low. Expansionary if debt is high

reason to believe the government's promises. As a response to this intrinsic lack of credibility, several countries have introduced rules in order to guide and constrain fiscal policy decisions.

This has been particularly vivid in Europe, where specific issues have surfaced in relation to the Economic and Monetary Union. The euro area has provided a rich laboratory for discussing and assessing fiscal policy effectiveness, decentralization versus centralization, and coordination.

In 2008–09, earlier views and established doctrines about the effectiveness (or lack of) of discretionary fiscal policy were revisited, with a wide agreement emerging on the usefulness of undertaking a substantial and coordinated fiscal expansion. This was largely based on the recognition that most of the

conditions that are required for fiscal policy to be effective were likely to be met:

- The world economy was hit by a major demand shock, resulting in a strongly negative global output gap and significant risks of deflation. So the aggregate supply curve could be expected to be flat.
- World long-term interest rates were very low and with near-zero policy rates, global monetary policy was strongly accommodative. So there was no crowding effect to talk of.
- The share of credit-constrained households and firms had increased as a result of lower bank willingness to lend. Hence, pouring public cash into them had a higher probability to raise demand than in normal times.
- The stimulus was coordinated or at least simultaneous the world over. So the ineffectiveness of fiscal expansion in a floating exchange-rate regime did not hold.

Nevertheless, when president-elect Barack Obama declared on 9 January 2009 that "there is no disagreement that we need action by our government, a recovery plan that will help to jumpstart the economy," dissenting voices were quick to make themselves heard.

3.3.1 Rules and principles for fiscal policy

In the 1990s and the 2000s, concerns about recurrent deficits and the sustainability of public debt led many governments to adopt budgetary rules. In principle, such rules aim at safeguarding sound government finance in a credible and sustainable way, while preserving the contribution of fiscal policy to contra-cyclical output stabilization. However, whether or not they succeed in achieving these objectives is a matter of design and enforcement. Good rules can improve policy, but bad ones can worsen it.

Rules play an important role in decentralized fiscal systems, in which the possibility of bailouts and the existence of transfers from central to sub-national governments may lead to excessive spending and inefficient resource allocation. Box 3.12 documents the US case.

Box 3.12 Fiscal Rules and Macroeconomic Stabilization in the United States

The US Constitution grants states a very large degree of fiscal autonomy, but sub-national (state and local) governments are subordinated to two kinds of fiscal rules (see the detailed description and discussion in Laubach, 2005). First, all US states except Vermont operate under balanced budget requirements. Second, more than half have adopted tax and expenditure limitations. State governments operate under fund accounting: All revenues accrue to, and all expenditure items are paid from specific funds.

Such funds typically include a general fund (operating budget) covering current revenues and expenditures, a capital fund, an insurance trust fund, a public employee retirement fund, and a budget-stabilization fund. Reserves accumulated during expansions can complement the stabilization role played by federal taxes and expenditures.

Balanced-budget requirements typically apply to the general fund (while capital spending can generally be financed by debt). They may take many forms, from a softer requirement to present a balanced budget to the state legislature (in 45 states) to a condition that the legislature passes a balanced budget (in 41 states) or that the governor may sign only a balanced budget (in 31 states). In 38 states, the budget has to be balanced at the end of the fiscal year, as there is a prohibition against carrying a deficit forward into the next fiscal year, enforced by a restriction on the issuance of general state debt.[35] There is empirical evidence that such requirements are effective in constraining states to adjust policies to keep current revenues and spending in balance. The price to pay, as signaled by Laubach (2005), is that they tend to induce pro-cyclical spending behavior, as states tend to cut core spending during downturns. This was apparent in 2009 when most US states would have been forced to cut spending programs, including social expenditures, in response to the recession. For this reason the stimulus program enacted in 2009 included federal transfers to state and local governments of the order of magnitude of 0.3% of GDP per year.

This suggests that budget-stabilization funds, when they exist, do not fully achieve their stabilization role. In some cases, tax and expenditure limitations may hamper the accumulation of reserves. In 35 states, stabilization funds are even capped at 10% or less of general fund expenditures. The limited counter-cyclical role of state budgets increases the responsibility of the federal government in responding to economic downturns.

Kopits and Symansky (1998) have identified eight criteria for an "ideal" fiscal rule:[36]

- a clear definition,
- transparent public accounts,
- simplicity,
- flexibility—in particular regarding the capacity to react to exogenous shocks,
- policy relevance in view of the objectives pursued,

35. These various schemes are not mutually exclusive.
36. See also Creel (2003).

- capacity of implementation with possibility of sanctioning nonobservance,
- consistency with the other objectives and rules of public policies,
- accompaniment by other effective policies.

This simple list suggests the existence of potentially delicate trade-offs, for example between simplicity and relevance, or between clarity and flexibility.

In practice, fiscal rules can be specified in various forms (table 3.7): Public debt ceilings; fiscal (financial or primary) deficit commitments; spending targets; assignment rules for fiscal surpluses; principles for the preparation of the budget. They can apply *ex ante* (to the budget submitted to vote) or *a posteriori* (to the observed results). In the EU, the Stability and Growth Pact,

Table 3.7
Examples of fiscal rules in force in the early 2000s

Rule/country	Enforcing date	Coverage[a]	Basic principle[b]	Escape clause[b]	Additional rule[b]	Statute[c]	Sanction[d]
Budget rules							
Argentina	2000	NG	OB/DL	CF	EL	L	J
Brazil	2001	NG, SG	CB		WL	L	J
Canada	Several	SG	CB			L	J
EU	1997	GG	OB/DL	MY		T	F
Germany	1969	NG, SG	CB			C	J
New Zealand	1994	GG	PB	MY		L	R
Peru	2000	NG	OB/DL	CF	EL	L	J
Switzerland	Several	SG	CB			L	J
US	Several	SG	CB	CF		C	J
Debt rules							
Brazil	2001	NG, SG	SL			L	J
Colombia	1997	SG	PL			L	J
EU	1997	GG	PL			T	J
New Zealand	1994	GG	SL			L	R

[a]General government (GG), national (central, federal) government (NG), subnational (including local) governments (SG).
[b]Budget rules consist of overall balance (OB), operating balance (PB, current income minus current expenditures including capital depreciation), or current balance (CB, current income minus current expenditures not including capital depreciation), subject to a prescribed limit on deficit (DL) as a proportion of GDP, applied on an annual basis, except if specified on a multi-year (MY) basis. A contingency fund (CF) is provided in some cases. Additional rules consist of limits on primary expenditure (EL) or wage bill (WL). Debt rules are specified as a limit for a given year (SL) or permanently (PL), as a proportion of GDP or of government revenue.
[c]Constitution (C), legal provision (L), or international treaty (T).
[d]Sanctions for noncompliance: Reputational (R), judicial (J), financial (F).

Source: Kopits (2001), table 1, p. 18.

described in greater detail below, combines an *ex post ceiling* for the deficit with a public debt reference target.

In some cases, the emphasis is put on fiscal balance. The balance requirement is frequently expressed with reference to the current budget (current spending has to be financed by current receipts): This is the so-called *golden rule of public finance**.[37] This rule, which was enshrined at the end of the 1960s in the German constitution (until it was reformed in 2009), authorizes use of debt to finance public capital expenditures only, unless there is a "disturbance of the macroeconomic equilibrium." The rationale is that debt finance better allows spreading out of the financing burden over the years during which the financed equipment will be productive, and that outstanding government debt is matched by (presumably profitable) government assets, which preserves government net wealth. Golden rules have supporters among economists (Blanchard and Giavazzi, 2004). However, a golden rule does not prevent debt from becoming unsustainable (if the counterpart of debt accumulation consists in assets of limited marketability, a government can become insolvent even though it has only borrowed for investment). Another problem is that the definition of public investment is open to criticism. A narrow definition tends to introduce a disputable bias in favor of brick and mortar spending at the expense of investment in human capital, but a broad definition may render the rule ineffective. Additionally, the focus on gross rather than net investment is disputable since only net investment benefits future generations. For these reasons, Germany replaced the golden rule in 2009 with a tighter rule whereby structural net borrowing is limited to 0.35% of GDP per annum starting in 2016 for the federal government, whereas the *Länder* will no longer be allowed to run any structural deficit starting in 2020. These limits on structural deficits may be violated only in exceptional circumstances such as natural disasters or severe economic crises. In such circumstances, the government would be required to provide an amortization plan to be approved by parliament. Under the new German rule, the cyclical component of the deficit also falls under close scrutiny, based on the same methodology as for the Stability and Growth Pact. Finally, any deviation by the implemented budget is recorded on a control account and must be netted out over time (see *Bundesfinanzministerium*, 2009).

In the late 1990s the UK government also adopted a fiscal policy framework based on two rules: The golden rule and a so-called "sustainable investment" rule assessed over the economic cycle. The rule worked well until the mid-2000s when it became clear, even before the crisis, that the commitment to

37. There is also a "modified golden rule," which includes the depreciation of public capital in current expenditures. It amounts to requiring that the growth in public debt does not exceed the net fixed-capital formation of the public sector. Note that this fiscal rule needs to be distinguished from the "golden rule" that characterizes balanced growth in the neo-classical growth model (chapter 6).

manage public finances over the cycle was difficult to monitor and enforce (see box 3.13).

Lastly, countries with a high public debt often choose rules targeted at the primary deficit. So did for example Belgium, which adopted a primary surplus floor of 6% of GDP at the end of the 1990s. The adjustment programs that the IMF imposes on countries in financial difficulty also include primary balance targets.

Box 3.13 The British "Golden Rule"

In 1998, a two-pronged fiscal rule was introduced in the UK. The *golden rule of public finance* only permits structural public deficits insofar as they have as a counterpart net public investment. The sustainable investment rule specifies that the ratio of net public debt to GDP has to remain at a "stable and prudent level" defined by the Chancellor of the Exchequer as no more than 40%. This latter rule applies *over the economic cycle*. Moreover, principles used for assessing private investments also apply to public investment decisions: A project shall be implemented only if the present value of expected returns covers the expenditure. This is intended to ensure that the debt incurred to finance investment projects does not jeopardize public finance sustainability.

This new approach notably aimed at protecting capital expenditures even in the face of strong fiscal restrictions. The underlying diagnosis was that current spending (notably on social security) had expanded to the detriment of net public investment.

The approach, however, raises practical difficulties:

- The net return on public investment is difficult to evaluate. When the infrastructure allows for the expectation of tolls (e.g. from motorways), forecasts of future receipts can be conducted on the basis of assumptions about frequency of use. When investment allows rationalizing public sector production (computerization of government, for example), the associated productivity gains can be estimated and quantified. The bulk of public investments, however, are there to meet new needs; not only do they generate no income nor savings, but very often they involve additional expenditures. This is typically the case with new construction projects, such as new hospitals or schools.
- The calculation of the net return on public investments can give rise to a problem of information asymmetry, whereby the proponents of investment projects may be tempted to over-estimate their return while the central government may not have all the necessary information to conduct a reality check. Also, the perimeter of public investment (gross fixed capital formation) excludes investment in

human capital, while the superiority of physical capital over human capital in terms of productivity has not been proven.
- The principle that debt stability applies over the economic cycle is economically sensible as it preserves the possibility of using budgetary policy for stabilization purposes, but cyclical corrections are technically questionable and can be easily manipulated.

In the UK, the implementation of the golden rule did allow a sharp recovery of net public investment which, as a percentage of GDP, had not ceased decreasing from the 1960s into the 1990s and had reached the extremely low level of 0.8% of GDP by 1996–97. Until 2007, this recovery of public investment was consistent with the fiscal balance staying within the Stability and Growth Pact boundaries. However, the deficit increased sharply in 2008–09 (figure B3.13.1). The debt ratio increased from 44% of GDP in 2007 to 72% in 2009, well above the level set by the rule. Although the crisis clearly had an exceptional character, this evolution highlighted the difficulty of delivering on a commitment to a given evolution "over the cycle" when both the length of the cycle and the magnitude of the fluctuations are unknown.

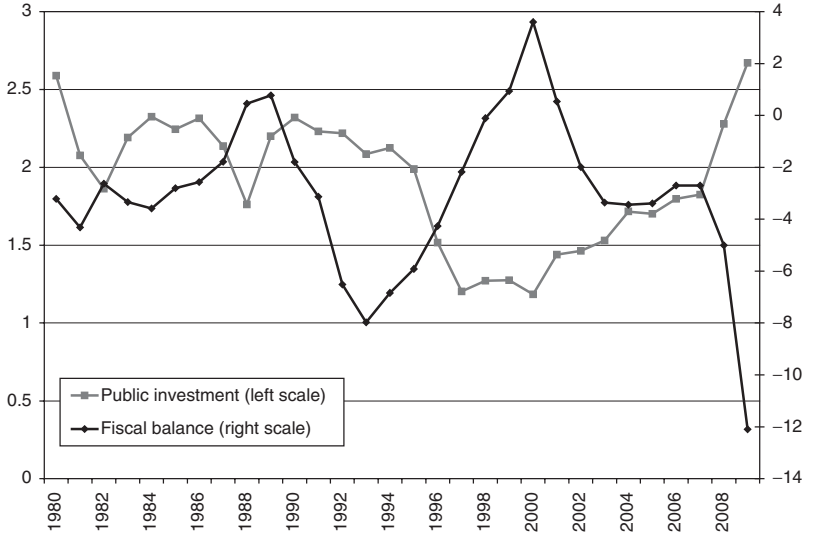


Figure B3.13.1 Public investment and fiscal balance in the UK (% of GDP). Sources: AMECO database, April 2009, and European Commission forecasts, Autumn 2009).

Beyond the choice of a specific rule, the very adoption of a fiscal rule raises questions: No rule is optimum in every circumstance. The golden rule, in particular, limits the capacity of governments to encourage

consumption smoothing over time.[38] It may also lead to an excessive level of public investment or skew public spending choices. Nevertheless, many governments, especially in Europe, value the disciplining character of a rule and are convinced that its advantages exceed its disadvantages, and have adopted such frameworks at national level. Experience suggests that their effectiveness is uneven as this depends on domestic political institutions. In particular, rules generally succeed better where governments are based on a coalition whose fiscal strategy is traditionally part of the coalition agreement (Hughes Hallett et al., 2001).

In the euro area, the fiscal rule (the Stability and Growth Pact, see below) can be viewed both as a shield against imprudent fiscal policies and as a way of minimizing the costs of coordinating policies among a large number of players, while ensuring equality of treatment among them. It may however have distracted attention from accumulating macroeconomic imbalances in some member states. Spain, especially, was hailed in the 2000s for impeccable budgetary discipline, but at the same time it let a real estate bubble of massive dimensions to develop and allowed a significant real exchange-rate appreciation.

The adoption of rules is not the only way of guiding governmental action. Another solution consists in reforming the institutions that are involved in the budgetary decision process. Political economy approaches have indeed shown that fiscal sustainability is affected by political and social conditions as well as by the quality of institutions:

- An unsustainable fiscal policy frequently results from conflicts between social groups on the division of the costs of a fiscal adjustment which is perceived as necessary, but the burden of which nobody wants to bear. In such a situation of fiscal *war of attrition**, each social group tends to delay the adjustment, hoping to shift the burden onto another group. The result is that fiscal adjustment is delayed.[39] The more polarized the political parties, the more frequent this kind of situation. The pension reform issue provides an example of such behavior: An agreement on the diagnosis does not automatically lead to reform if social groups and/or political parties disagree on the sharing of the burden between capital and labor incomes. *A contrario*, when there is a trans-partisan agreement, like in the US (with the 1983 Greenspan Commission on retirement reform) or in Sweden (with the 1994 reform prepared by an agreement between the main political parties), the adjustment can be quickly undertaken. From a more general perspective, social and political fragmentation (between social groups, generations, regions, etc.) weakens solidarity and undermines fiscal discipline, because each

38. See Buiter (1998).
39. See the formal approach developed by Alesina and Drazen (1991).

group focuses on its own interests and tends to neglect the collective costs of an unsustainable fiscal policy.[40]

- The quality of budgetary institutions and procedures explains an important part of the performance divergence across countries. When the budgetary decision is only lightly centralized, the multiplicity of demands on public finance leads to a defective control of the deficit. Contrarily, empirical analyses confirm that deficits are better kept under check when a single authority monitors the preparation of the budget, when the government is in a position to reject parliamentary amendments that increase the deficit, and when the Ministry of Finance controls the implementation of the budget. More generally, the degree of fiscal centralization accounts for country divergences on fiscal deficit and debt (von Hagen and Harden, 1994). In France, for instance, the health insurance budget is voted by Parliament but it is not binding. The lack of control over implementation results in a systematic overshooting of the objectives. In contrast, the French Minister for Budget has very strong powers over central government spending. The central budget may be voted in deficit, but its implementation stays close to the voted figure.

Going further, it would be conceivable to borrow from the institutional setup of monetary policy (cf. chapter 4) by entrusting an independent agent with the responsibility for setting the annual fiscal balance objectives to be respected. Charles Wyplosz (2002) has suggested, for example, that an independent fiscal policy committee be in charge of fixing each country's yearly fiscal balance targets, leaving to governments the choice of the fiscal instruments to meet the targets. The yearly fiscal balance targets would be fixed in advance of the budget preparation process and would have force of law. The committee would also approve the draft finance law.

This proposal was not taken up by EU governments and is unlikely to be implemented any time soon. It illustrates, however, a renewal of the intellectual approach to fiscal policy and it has received an echo, albeit in a reduced form, in the Sapir report (2004) prepared for the President of the European Commission. The report recommended putting in place, in each country, an independent budget audit committee, which would have no decision-making capacity but would have access to all relevant data and would publish its assessments.

3.3.2 Fiscal policy in the European Monetary Union

Three major issues have surfaced in the fiscal policy debate in the context of the European Economic and Monetary Union (EMU), which deserve specific discussion: Fiscal discipline, fiscal federalism, and fiscal policy coordination.

40. The French pamphleteer Frédéric Bastiat thus defined Government as "the great fiction through which everybody endeavors to live at the expense of everybody else." (Bastiat, 1848).

a) Fiscal discipline

The main argument for fiscal discipline in a monetary union is based on the risk that an unsustainable fiscal policy in a member state would endanger monetary stability in the whole area. Suppose that a government, after years of fiscal profligacy, is now on the verge of a solvency crisis. Its bonds are charged a high risk premium by investors.[41] The possibility of a funding crisis affecting a euro area country whose public finances are perceived to be weak was proven not to be a pure fantasy. The European Central Bank could then be subject to political pressures to acquire some bonds directly and to monetize them, which could jeopardize the central bank's objective of a low inflation rate. To prevent such risks, the Maastricht Treaty precludes public securities being directly purchased by euro area central banks (which is of little effect as it does not prohibit buying them on the secondary market). However, the risk does not stop there. The government in difficulty would tend to borrow through short-term instruments (because no private investor would agree to lend it over the long run, for fear of default), and the central bank would soon face a dilemma: Either bring the government to the brink of a failure through nonaccommodating monetary policy, or cut rates to preserve the government's capacity to pay. Another line of argument emphasizes systemic risk in an integrated financial market: For example, if a state's debt is held by banks located in other euro area countries, a default on its debt would weaken the whole area's financial sector (Eichengreen and Wyplosz, 1998).

This is the rationale for fiscal discipline in a monetary union.[42] However, while most economists would agree on it, the debate remains lively on the appropriate procedures to enforce such discipline. In principle, financial markets should be able to price the risk of sovereign default and exercise pressure on governments whose sustainability is uncertain. In practice however, markets may underprice risk for protracted periods, failing to incentivize public finance adjustment. Furthermore, in the euro area, markets may anticipate the bail-out of a country in difficulty by its partners, which would result in lower risk premia. For these reasons the avoidance of "excessive deficits" and the possibility of sanctions against offenders are enshrined in the EU treaty. In the run-up to monetary union, Germany insisted on an enforcement mechanism, which gave rise to the *Stability and Growth Pact** (hereafter SGP, see box 3.14) of 1997.

The SGP aims at enforcing fiscal discipline while leaving some room for counter-cyclical policy. In accordance with the treaty, the SGP requires that EU member states keep their public deficits and debt levels below 3% and 60% of domestic GDP, respectively, and provides for financial sanctions when the

41. The scenario could have been regarded as overdone until the bond spreads across euro area members started to widen at the end of 2008, reaching 300 basis points in the first quarter of 2009 (see figure B3.1.1).

42. For a detailed discussion, see Buti et al. (2003), Pisani-Ferry (2003), and Coeuré and Pisani-Ferry (2006).

deficit exceeds the 3% ceiling. In its original form, the SGP was extensively criticized by the economic profession on a number of grounds:

- An excessive focus on short-term considerations at the expense of long-term analysis, since it put emphasis on the deficit rather than on debt.
- Asymmetry and pro-cyclical bias, since it incentivized participating countries' governments to reduce their deficit in bad times (in order not to breach the 3% threshold) and not in good times.
- Lack of economic underpinning, as no theory validates the long-term target of the zero debt-to-GDP ratio implicit in the SGP's initial call for budgets close to balance or in surplus. On the contrary, it is legitimate to entertain some debt to finance public investment as long as the social return of the latter exceeds the cost of the former.
- A one-size-fits-all approach, even though states differ in their initial situations (e.g., their debt and public asset levels) as well as in their long-term prospects (long-term growth, inflation, and off-balance liabilities).
- Weak enforcement, as sanctions carry very limited credibility.

Box 3.14 The Stability and Growth Pact

During the negotiation of the European Economic and Monetary Union, in the early 1990s, it was agreed that member states should avoid "excessive deficits" (Article 104 of the Maastricht Treaty) and should face sanctions if this discipline went unobserved. Reference thresholds of 3% of GDP for the deficit of the general government and 60% of GDP for gross public debt were agreed upon on this occasion and were laid down in the protocol on the excessive deficit procedure annexed to the Treaty, together with the possibility of sanctions against delinquent countries. However, the Treaty did not specify the procedure for implementing those sanctions.

There is no clear rationale for the 3% and 60% figures but there is some consistency between them: A maximum deficit of 3% of GDP ensures the stability of a public debt ratio of 60% of GDP when GDP increases by 5% a year in current euros, which corresponds to an inflation rate of 2% a year (European Central Bank (ECB) ceiling) and a rate of real growth of 3% (potential output growth at the time of treaty negotiations), cf. box 3.7. Ideally, the deficit threshold should have been differentiated according to the growth potential of member states, some among them having higher growth prospects due to lower initial GDP per capita. But the need prevailed for a simple, across-the-board rule that would facilitate the political discussion and provide markets with a credible fiscal discipline commitment.

On the eve of the introduction of the euro, the German government demanded that the fiscal discipline commitments and the procedures for sanctioning undisciplined member states be detailed and conveyed in an enforceable document. The Pact comprises two main elements:

- *A preventive arm.* Each member state is to adopt a medium-term objective for its cyclically adjusted budgetary position that is consistent with the overall objective of being close to balance or in surplus and leaves room for stabilization of normal cyclical fluctuations without breaching the 3% threshold. It prepares a three-year *stability program** (in non-euro-area countries aspiring to become members, it is called a *convergence program**) that is updated every year and is submitted to the assessment of the Commission and to the approval of the Council of Finance Ministers (*Ecofin**). The program describes the adjustment path toward the medium-term objective, taking as a benchmark a 0.5 percentage point improvement in cyclically adjusted terms per year. Initially the focus was on headline deficits but over time the EU has gradually moved toward monitoring cyclically adjusted deficits.
- *A dissuasive arm.* Except if "exceptional and temporary," the headline (financial, i.e. non-cyclically adjusted) fiscal deficits of member states should never exceed 3% of GDP. Initially, the Pact defined as "exceptional" a year during which real GDP falls by at least 2%, but this threshold was revised to 0 in 2005. When the deficit threatens to reach, or exceeds, the 3% threshold, a specific surveillance procedure (the *excessive deficit procedure*) is set in motion according to a predetermined timetable of increasing pressures: Steps include early warning, identification of an excessive deficit, recommendation to implement corrective actions, obligation to make a non-interest-bearing deposit with the Commission, conversion of the deposit into a fine. These various stages, in particular sanctions, give rise to decisions by the euro area finance ministers under a qualified majority vote (i.e., with the voting weights usually applied to the member states in the European decisions). The fine includes a fixed component of 0.2% of GDP and a variable component linked to the size of the deficit (0.1% of GDP per percentage point in excess of the 3% limit), within an annual limit of 0.5%, but no fine has ever been considered in practice.

After extensive criticism of the SGP, the failure of several member states to comply with it and subsequent 2003 decision by euro area finance ministers to put the pact "in abeyance" instead of simultaneously activating its corrective procedures against France and Germany, a substantial reform

of the SGP was adopted in 2005.[43] While the 3% and 60% thresholds for the deficit and the debt were kept unchanged, the reform introduced significant flexibility in order to "enhance the economic rationale of the budgetary rules to improve their credibility and ownership." In addition to the emphasis on cyclically adjusted figures, the medium-term budgetary objective (MTO) of "close to balance or in surplus" was replaced by individual MTOs that recognize the specific economic characteristics, situations, and structural reform objectives of each member state. Implicit public liabilities such as pensions are also taken into account in the assessment of the budgetary situation, as well as systemic pension reforms that may lead to a short-term deterioration of the deficit but improve the longer term sustainability of public finance. Moreover, at German insistence it was agreed that "other relevant factors" are to be taken into account when estimating whether a member state complies with budgetary discipline—clearly a potential loophole.

With the widening of budgetary deficits in 2008–09, the vast majority of EU member states found themselves in breach of the no-excessive-deficit provision. As regards the exceptional-circumstances clause, the European Commission considered that "although the excess [deficit] over the reference value can be regarded as exceptional it is not temporary in the sense of the Treaty and the Stability and Growth Pact" (Article 104(3) reports of the Commission, 7 October 2009).

Proposals for reforms included:

- *Different targets*: More focus on the debt rather than on the deficit, and on the cyclically adjusted deficit rather than on the headline deficit; account for off-balance liabilities (Buiter and Grafe, 2003; Coeuré and Pisani-Ferry, 2006); exclusion of some capital expenditures from the deficit to be monitored (Blanchard and Giavazzi, 2004).
- *Different threshold values*: It was suggested that member states be differentiated depending on their long-term growth rate and on the initial level of their debt.
- *Better incentives over the cycle*: More deficit reduction in good times, while allowing countries to exceed the SGP deficit ceiling during economic slowdowns (Buti et al., 2003).

The 2005 SGP reform addressed several of these issues. In particular, the medium-term fiscal objectives are no longer the same for all member states; instead they vary according to the country's potential growth rate, debt level, and implicit liabilities. More time to adjust is left to countries

43. See the *Presidency Conclusions* of the 22–23 March, 2005 European Council, Annex II.

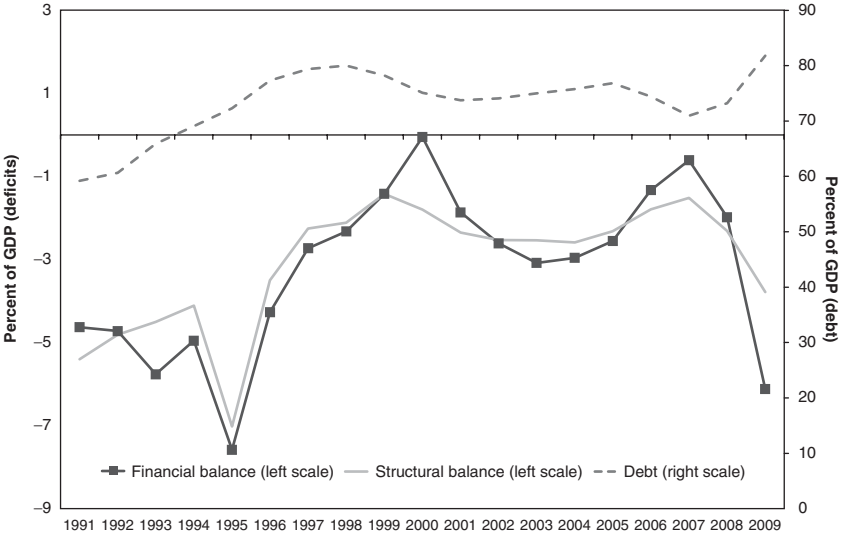


Figure 3.19 Indicators of budgetary discipline in the euro area, 1991–2009.
Source: OECD Economic Outlook 86, November 2009.

experiencing negative growth or undergoing structural reforms that enhance debt sustainability in the long run (such as pension reforms). Finally, the cyclically adjusted deficit plays a central role in the so-called preventive arm of the SGP (see box 3.14).

Some have seen this reform of the SGP as a watering down that deprives the SGP of its teeth (Calmfors, 2005). Its implementation nonetheless seems to have promoted some discipline, as suggested by figure 3.19:[44] On the whole, deficits have decreased in the euro area after enactment of the SGP reform in 2005. Like in other regions, the 2007–09 crisis resulted in a sharp deterioration of public finances in the EU. Furthermore, some countries, such as Ireland and Spain, that were considered highly disciplined before the crisis due to balanced budgets or even financial surpluses, abruptly turned to deep deficits. This crisis has highlighted the limitations of the SGP for monitoring discipline. Indeed the focus on public finance had the unfortunate consequence of making policymakers blind to the large imbalances being built up within the euro area.

b) Fiscal federalism in Europe

In federal states, the macroeconomic stabilization function is generally assigned to the federal level, while individual states are often subject,

44. Manasse (2007) argues that the fiscal indiscipline in large European countries during recession years does not imply that SGP rules are ineffective. In his view, fiscal deficits would have been even higher in the absence of such rules.

sometimes by their own will, to strict fiscal discipline. In the US, when a state undergoes a negative income shock, its residents pay lower federal taxes but continue benefiting from federal spending (on public goods, transfers, etc). The federal budget therefore functions as an automatic shock absorber. Sachs and Sala-i-Martín (1992) have found that up to 30–40% of economic shocks that affect the states are absorbed by the federal budget. This evaluation has been debated and the current consensus is rather around 20% (Melitz and Zumer, 2002), which is still not negligible.[45]

In the euro area, the choice not to supplement the single currency with a federal budget was made in the early 1990s for political reasons: Monetary union already was a step toward a European federation, and that was the limit of what governments and public opinion could accept. Furthermore, as public spending is already high, this would have required transfer of budgetary functions from the national to the European level.

The European budget (see box 3.15) could play a stabilizing role only if its relative importance increased and if its spending and income were more sensitive to the business cycle. This would require a major change of political organization that might for example consist in transferring major social security functions such as unemployment insurance to the EU level, or, in the absence of a federal budget, in creating an automatic compensation mechanism through the European budget for variations in national fiscal receipts (Italianer and Pisani-Ferry, 1992). In the absence of an improbable large-scale reform, the stabilization function falls therefore on national budgets, which raises the question of policy coordination.

Box 3.15 What Is the EU Budget Used for?

In 2010, the European Union (EU) budget amounted to 122.9 billion euros, corresponding to 1.04% of EU gross national income.[46] In contrast, national budgets represent from 40 to 60% of the member states' income. The economic policy responsibilities at the Union level are limited (see chapter 2). Indeed, its budget only fulfils interregional redistribution and allocation functions, and even that only in addition to national policies. Since the budget cannot be in deficit, spending is limited by available resources. Although the share of agriculture has sharply declined since the early 1980s, EU expenditures are still heavily concentrated on the *common agricultural policy** (CAP) and rural development (42% of total expenditures in 2010) and cohesion (convergence, regional development, etc., 45%).

45. Bayoumi and Masson (1998) find that the Canadian federal government contributes to stabilizing 17% of shocks faced by the provinces.

46. Payment appropriations figure. After tough negotiations, EU member states agreed in December 2005 on financial perspectives for 2007–13 with the EU budget fixed at 1.045% of gross national income.

The Commission handles expenditures and, when they are delegated to states or to local authorities (as is the case for farm spending and for structural funds), monitors fund use. In the event of irregularity, an inquiry can be conducted by the European Anti-Fraud Office.

There is no European tax. The budget is financed out of member states' contributions based on their gross national income (76% of resources in 2010), on VAT receipts transferred to the EU (half a percentage point, producing 11% of the budget), and, finally, on custom duties and levies on agricultural imports (12%).

As argued by Sapir et al. (2004), the structure of the budget poorly distinguishes the allocation, redistribution, and stabilization functions. The CAP, originally intended to ensure Europe's food safety and to increase agricultural productivity, increasingly looks like a redistribution policy for farmers. This confusion of objectives causes inefficiency and tensions between the member states.

A number of other factors may over time lead to the development of the European budget. In ever more integrated markets, an increasing number of functions belonging to the state (safety, consumer protection, regulation of markets) are now implemented at the Community level. In some tightly integrated sectors or in sectors exhibiting a natural transnational dimension (for example, transport), infrastructure investment is a true European public good and it is easy to imagine that it could be financed at the European level. Lastly, "European" taxes, such as green taxes, could emerge as a way to finance the pursuit of common objectives. However, even an unlikely quadrupling of the EU budget would not transform it into a significant macroeconomic instrument.

c) Fiscal policy coordination

The economic literature traditionally identifies two major reasons for nations to coordinate economic policies. The first is the provision of the international public goods that decentralized action will in general fail to produce. The second relates to the sub-optimality of uncoordinated decisions in the presence of externalities, even for the pursuit of predominantly national objectives.

These two reasons apply in Europe and especially in the euro area. First, safeguarding the single market and its proper functioning, as well as financial stability, can be viewed as an EU-wide public good. This justifies various forms of coordination, including mutual recognition, harmonization of some regulations and taxes, or EU-wide competition and bank-supervision policies.

Second, the advent of European Monetary Union has introduced specific fiscal policy externalities (cf. box 3.16). In a monetary union, an expansionary fiscal policy in one country creates a positive demand-externality for the other members but—if the central bank responds by raising the interest rate—a negative interest-rate externality. This is a second, different justification for mutual surveillance of national fiscal policies.

Moreover, a number of political economy arguments also call for coordination. Coordination may strengthen the credibility of the national fiscal plans (which rest on a number of hypotheses that may not be common but are at least discussed jointly), and peer pressure facilitates their implementation by reducing the ability of parochial interest groups to successfully divert policies from the pursued objectives. Simultaneously, in a single monetary area, policy coordination between governments gives them a collective responsibility which may alleviate the risk that public opinion might see the central bank as the sole institution responsible for economic policy in the zone.

Box 3.16 Fiscal Policy Spillovers within a Monetary Union

Suppose two identical countries called A and B form a monetary union (U), and let us represent the equilibrium in the IS–LM diagram as in box 3.5. The three panels in figure B3.16.1 represent equilibrium respectively in country A, in country B, and in the monetary union U.

Now let us assume that both countries face a negative, symmetric demand shock, for example, a fall in imports from a third country that buys goods from A and B. In the absence of an economic policy reaction, the IS curve of each country, and therefore also the IS curve for the whole union, moves to the left. The fall in output in each country is limited by the fall in the interest rate (new equilibrium at E′ in panel *U*). If country A reacts to the demand shock by an expansionary fiscal policy, this policy brings its IS curve back to the right. The IS curve of country B also moves (but to a lesser extent) toward the right, because it benefits from increased exports to country A. The aggregate IS curve for the whole monetary union also returns partially toward the right. The fall in the interest rate is less, not only for country A, but also for country B. The latter profits from increased exports to country A, but suffers from a lesser fall in interest rates (equilibrium at E″).

If the interest rate externality dominates, country B is also likely to react by an expansionary fiscal policy. Output will then be stabilized, but at the price of a higher budget deficit, while a fall in the interest rate would have benefited the country's public finance.

In Europe, the trade channel seems to dominate the interest-rate one, so that a fiscal expansion in one country raises demand and output in

other countries, although the spillover seems to be limited and concentrated on neighboring countries (see Bénassy-Quéré and Cimadomo, 2006). This means that, short of policy coordination, each country feels little incentive to implement a stabilizing fiscal expansion, since part of the expansion benefits other countries.

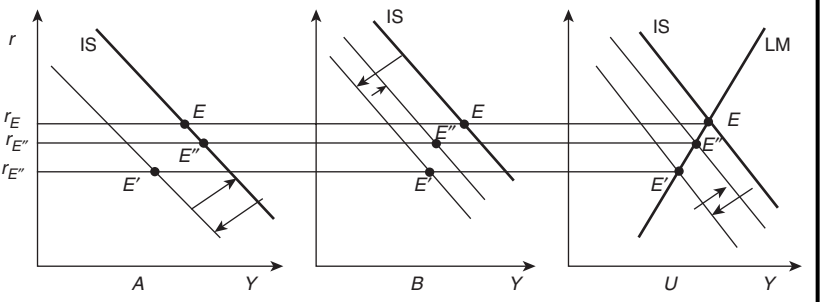


Figure B3.16.1 External effects of the budgetary policy in an IS–LM model.

Economic theory, however, also provides arguments against coordination. As discussed above, the Mundell–Fleming model, under flexible exchange rates and perfect capital mobility (assumptions adapted to the situation of the euro area as a whole) leads to the conclusion that fiscal policy is ineffective for the Union as a whole;[47] the task of stabilizing the economy of the euro area accordingly falls primarily on monetary policy. The coordination of fiscal policies to stabilize the activity appears to be a second-best solution ranking below monetary policy, to be used, for example, when the ECB is prevented by a conflict among objectives from providing a needed stimulus (cf. chapter 4) or when monetary policy alone cannot stabilize the economy, as in 2008–09. Conversely, the Mundell–Fleming model also underlines fiscal policy effectiveness under fixed exchanges with perfect capital mobility, a situation that characterizes each member country in relation to its partners within the area. This discussion leads to the following policy assignment rule:

- The common monetary policy responds to shocks affecting the whole area (symmetric shocks).
- National fiscal policies respond to country specific shocks or shocks affecting a group of countries in relation to others within the area (asymmetric shocks).

47. Actually, even under the assumptions of the model, the inefficiency of fiscal policy holds only for a small country, which does not describe the euro area as compared to the rest of the world. For an open economy under flexible exchange rates, however, fiscal stimulus results in stimulating the trading partners' economies.

Finally, the central issue in Europe is not so much the coordination of fiscal policies *per se* as it is the nature of the overall policy mix, involving monetary policy. A general result from the analysis of second-rank optima addressed in chapter 1 suggests that, however desirable global coordination may be, partial coordination, for example, between fiscal policies only (with an uncooperative central bank) does not necessarily improve the situation nor bring the optimum any closer. Using fiscal policy coordination to counter the action of the ECB could only result in a costly failure. However, there is the risk that calling for coordination between fiscal and monetary policies might challenge the independence of the ECB. Indeed, some argue that coordination is unnecessary as long as national policies and the single monetary policy are specified correctly and every country is allowed to correct its own faults. Such was the German government's position at the time of the drafting of the Maastricht Treaty, which explains why economic policy coordination among member states is described in detail (Article 99) but also why coordination between governments and the ECB is only very succinctly addressed.[48] This position is very widely shared among European economists and governments (for an illustration, see Alesina et al., 2001).

This discussion suggests that in normal times, the objectives of fiscal policy coordination within the euro area should be carefully specified: First, ensure that national fiscal policies can play their stabilizing role at the local level, which requires managing fiscal deficits across the business cycle under a sustainability constraint (which brings us back to the role of fiscal rules and to the SGP); second, create the conditions for a dialogue between fiscal and monetary authorities about the economic diagnosis and the suitable responses (this is essentially the role of the Eurogroup); finally, ensure that when monetary policy alone cannot reach the objectives, it is possible to elaborate and implement common fiscal policy guidelines.

The deep recession of 2008–09 provided a textbook case for fiscal policy coordination within the euro area (and beyond):

- The shock was largely symmetric, implying that a common response was in order;
- Monetary policy was rendered ineffective by the state of the banking system, which called for additional budgetary support;
- The central banks were pursuing an accommodative monetary policy, reducing the policy rate to zero.

However, the fiscal response in Europe was only loosely coordinated. Although the European Council endorsed the Commission proposal for

48. It amounts in practice to the possibility, for the minister who chairs the ECOFIN Council, of attending the ECB governing council without voting, and to the rather vague requirement set out in the Treaty that " . . . without prejudice to the objective of price stability, the European System of Central Banks shall support the general economic policies in the Community" (see also chapter 4, box 4.6).

a coordinated stimulus package at the end of 2008, in the absence of an effective coordination mechanism, national responses varied considerably in size and composition. Small countries and countries whose underlying fiscal situation was weak provided a lower stimulus, while in some countries the stimulus was biased toward domestic industries.

3.3.3 Discretionary fiscal policy in times of crisis

Taylor (2000) formulated what he saw as a widespread agreement among economists about the policy implications of decades of accumulated theoretical and empirical research on fiscal policy: "In the current context of the US economy, it seems best to let fiscal policy have its main countercyclical impact through the automatic stabilizers." He also argued that discretionary fiscal policy should be "saved" for longer-term issues.

Martin Feldstein (2002) highlighted three reasons why a general consensus against discretionary fiscal policy had emerged in the US, all pointing to issues that we addressed in section 3.2. First, the old Keynesian view of high fiscal multipliers has been challenged by both theoretical approaches and empirical studies. Second, in some instances, fiscal policy can have anti-Keynesian impacts. Third, well-intentioned fiscal policy can be destabilizing due to policy lags and uncertainty about the economic response to fiscal changes. In his view, "there is now widespread agreement in the economics profession that deliberate 'countercyclical' discretionary fiscal policy has not contributed to economic stability and may have actually been destabilizing at particular times in the past."

However, he went on arguing that fiscal policy (preferably based on tax reductions rather than increases in spending) could be effective (and preferable to overly lax monetary policy) in situations characterized by low demand, low inflation, and low interest rates.

This is exactly the situation the US and many other countries found themselves in during the crisis of 2007–09. Feldstein (2009) argued that this downturn differed from previous recessions in that it was not due to high interest rates and could not be fixed by a reversal of monetary policy. Interest rates were reduced dramatically, but dysfunctional credit markets prevented the transmission of low interest rates to the economy. By the end of 2007, in the wake of the subprime crisis, Feldstein and others advocated a fiscal stimulus through temporary tax rebates. However, most of the $80 billion tax rebate passed by Congress in early 2008 was saved and consumer spending responded only weakly. Feldstein (2009) concluded in favor of a fiscal package based on increased government spending. In his view, the traditional problems associated with the use of expansionary fiscal policies were less present in 2008–09: Very easy money and dysfunctional credit markets would mean that crowding out through higher interest rates would not occur; the probable duration of the downturn would also limit the problem of time lags and spending delays.

The recourse to active and massive fiscal policy, however, further strengthens the necessity to put in place robust medium-term fiscal frameworks and to promote structural reforms to boost potential growth, as advocated for instance by Corsetti et al. (2009) who show that stimulation packages are more efficient when followed by spending reversals. The reason is that private consumption will react positively to a temporary increase in public spending only if households do not anticipate a permanent deterioration of fiscal balances.

In conclusion, decades of work on fiscal policy have not produced a universal, atemporal doctrine of use, nor do they lead to a "one-size fits all" set of recommendations. Instead, they have provided us with an analytical toolbox and a wealth of empirical studies that are particularly relevant to inform policymaking not only in normal times, but also in historical times such as the 2008–09 economic crisis. This is the best contribution to avoiding some of the mistakes made in the past, and in mitigating effectively both inflation and deflation, a problem that had already received Keynes's attention.

References

Alesina, A., and A. Drazen (1991), "Why are Stabilizations Delayed?," *American Economic Review*, 81, pp. 1170–88.

Alesina, A., and R. Perotti (1995), "Fiscal Expansions and Adjustments in OECD Countries," *Economic Policy*, 21, pp. 207–48.

Alesina, A., J. Galí, F. Giavazzi, H. Uhlig, and O. Blanchard (2001), "Defining a Macroeconomic Framework for the Euro Area," *Monitoring the European Central Bank*, no. 3, CEPR.

Auerbach, A.J., W. Gale, and P. Orszag (2004), "Sources of the Long-Term Fiscal Gap," *Tax Notes,* May (available at www.brookings.edu).

Auerbach, A.J., J. Gokhale, and L.J. Kotlikoff, (1991), "Generational Accounting: A New Approach to Understanding the Effects of Fiscal Policy on Saving," *The Scandinavian Journal of Economics*, 94, pp. 303–18.

Barro, R. (1974), "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November–December, pp. 1095–117.

Barro, R. (1979), "On the Determination of Public Debt", *Journal of Political Economy*, October, pp. 940–71.

Bastiat, F. (1848), "L'Etat", *Journal des Débats*, 25 September, English translation under the title "Government" available on www.bastiat.org.

Bayoumi, T., and P. Masson (1998), "Fiscal Flows in the United States and Canada: Lessons for Monetary Union in Europe," *European Economic Review*, 39, pp. 253–74.

Bayoumi, T., and S. Sgherri (2006), "Mr. Ricardo's Great Adventure: Estimating Fiscal Multipliers in a Truly Intertemporal Model," IMF working paper 2006-168.

Bénassy-Quéré, A., and J. Cimadomo (2006), "Changing Patterns of Domestic and Cross-Border Fiscal Policy Multipliers in Europe and the US," CEPII working paper 2006-24, December.

Bertola, G., and A. Drazen (1993), "Trigger Points and Budget Cuts—Explaining the Effects of Fiscal Austerity," *American Economic Review*, 83, pp. 11–26.
Bismut, C., and P. Jacquet (1997), "Fiscal Consolidation in Europe," *Cahiers de l'IFRI*, no. 18, Institut français des relations internationales.
Blanchard, O. (1985), "Deficits, Debts, and Final Horizons," *Journal of Political Economy*, 93, pp. 223–47.
Blanchard, O. (1993), "Suggestions for a New Set of Fiscal Indicators," in Vergon H., and F. van Winden (eds.), *The Political Economy of Government Debt*, Elsevier Science Publishers, chapter 14, pp. 307–25.
Blanchard, O. (2005), *Macroeconomics*, Prentice Hall.
Blanchard, O., and F. Giavazzi (2004), "Improving the SGP through a Proper Accounting of Public Investment," CEPR discussion paper, no. 4220.
Blanchard, O., and R. Perotti (2002), "An Empirical Characterisation of the Effects of Changes in Government Spending and Taxes on Output," *Quarterly Journal of Economics*, 117, pp. 1329–68.
Blanchard, O., J.-Cl. Chouraqui, R. Hagemann, and N. Startor (1991), "The Sustainability of Fiscal Policy: New Answers to an Old Question," NBER working paper No. 1547.
Boissinot, J., C. L'Angevin, and B. Monfort (2004), "Public Debt Sustainability: Some Results on the French Case," working paper G 2004/10, Paris: Institut National de la Statistique et des Etudes Economiques (INSEE).
Boorman, J. (2002), "Sovereign Debt Restructuring: Where Stands the Debate?," speech, 17 October (available on the IMF Web site).
Briotti, M.G. (2005), "Economic Reactions to Public Finance Consolidation: A Survey of the Literature," Occasional papers series No 38, European Central Bank.
Buiter, W. (1985), "A Guide to Public Sector Debt and Deficits," *Economic Policy*, no. 1, pp. 14–79.
Buiter, W. (1990), *Principles of Budgetary and Financial Policy*, MIT Press.
Buiter, W. (1998), "Notes on *A Code for Fiscal Stability*," NBER working paper no. 6522, April.
Buiter, W., and C. Grafe (2003), "Patching up the Pact: Some Suggestions for Enhancing Fiscal Sustainability and Macroeconomic Stability in an Enlarged European Union," *Economics of Transition*, 12, pp. 67–102.
Bulow, J., and K. Rogoff (1989), "Sovereign Debt: Is to Forgive to Forget ?," *American Economic Review*, 79, pp. 43–50.
Bundesfinanzministerium (2009), "Reforming the Constitutional Budget rules in Germany," Public Finance and Economic Affairs Directorate, Deficit Rule Reform Team, August.
Buti, M., and A. Sapir (1998), *Economic Policy in EMU*, Clarendon Press.
Buti, M., S. Eijffinger, and D. Franco (2003), "Revisiting the Stability and Growth Pact: Grand Design or Internal Adjustment?," *European Economy*, Economic Papers, 180.
Calmfors, L. (2005), "What Remains of the Stability and Growth Pact and What's Next?," SIEPS study no. 2005: 8, SIEPS: Stockholm.
Carroll, L. (1889), *Sylvie and Bruno*, re-edited, Indypublish, 2003.
Coeuré, B., and J. Pisani-Ferry (2006), "Fiscal policy in EMU: Towards a sustainability and growth pact ?," *Oxford Review for Economic Policy*, 21, 598–617.
Corsetti, G., A. Meier, and G. Müller (2009), "Fiscal Stimulus with Spending Reversals", IMF working paper, 09/106, May.

Cour, P., E. Dubois, S. Mahfouz, and J. Pisani-Ferry (1996), "The Cost of Fiscal Retrenchment Revisited: How Strong is the Evidence?," CEPII working paper 1996–16, December.

Creel, J. (2003), "Ranking Fiscal Rules: The Golden Rule of Public Finance vs. The Stability and Growth Pact," OFCE working paper, No 2003–04.

De Mello, L., P.-M. Kongsrud, and R. Price (2004), "Savings Behaviour and the Effectiveness of Fiscal Policy," OECD working paper ECO/WKP No 20, July.

Eichengreen, B., and Ch. Wyplosz (1998), "The Stability Pact: More than a Minor Nuisance?," *Economic Policy*, 13, pp. 65–104.

Elmendorf, D., and N.G. Mankiw (1999), "Government Debt," in Taylor J., and M. Woodford (eds.), *Handbook of Macroeconomics*, vol. 1C, North Holland, pp. 1615–69.

European Commission (2009), *Sustainability Report 2009*, September.

Executive Office of the President and Council of Economic Advisers (2009), "The Economic Case for Health Care Reform: Update", 15 December 2009.

Feldstein, M. (2002), "The Role for Discretionary Fiscal Policy in a Low Interest Rate Environment", NBER working paper 9203, Cambridge: National Bureau of Economic Research.

Feldstein, M. (2009), "Rethinking the Role of Fiscal Policy", NBER working paper 14684, Cambridge: National Bureau of Economic Research.

Fleming, J.M. (1962), "Domestic Financial Policies Under Fixed and Floating Exchange Rates," *IMF Staff Papers* 9, 369–80.

Giavazzi, F., T. Jappelli, M. Pagano, and M. Benedetti (2005), "Searching for Non-Monotonic Effects of Fiscal Policy: New Evidence," *Monetary and Economic Studies*, Institute for Monetary and Economic Studies, Bank of Japan, 23, pp. 197–217.

Hamilton, J., and M. Flavin (1986), "On the Limitation of Government Borrowing: A Framework for Empirical Testing," *American Economic Review*, 76, pp. 808–19.

Hansen, A.H. (1953), *A Guide to Keynes*, McGraw-Hill.

Heller, W. (1966), *New Dimensions of Political Economy*, Harvard University Press.

Hemming, R., M. Kell, and S. Mahfouz (2002), "The Effectiveness of Fiscal Policy in Stimulating Economic Activity – A Review of the Literature," IMF working paper WP/02/208.

Hicks, J. (1937), "Mr. Keynes and the "Classics": A Suggested Interpretation." *Econometrica*, 5, pp. 147–59.

Hughes Hallett, A., R. Strauch, and J. von Hagen (2001), "Budgetary Consolidations in EMU", European Commission Economic Papers 148, Brussels: European Commission.

International Monetary Fund (2003), *IMF Concludes 2003 Article IV Consultation with the United States*, 5 August (available on the IMF Web site).

International Monetary Fund (2009), *The State of Public Finances: Outlook and Medium-Term Policies After the 2008 Crisis*, Fiscal Affairs Department, Washington D.C.: International Monetary Fund.

Italianer, A., and J. Pisani-Ferry (1992), "Systèmes budgétaires et amortissement des chocs régionaux", *Économie Prospective Internationale*, no. 51, Summer.

Joumard, I., M. Minegishi, C. André, C. Nicq, and R. Price (2008), "Accounting for One-off Operations when Assessing Underlying Fiscal Positions," OECD Economics Department working papers no. 642, Paris: Organisation for Economic Cooperation and Development.

Keynes, J.M. (1931), *Essays in Persuasion*, Harcourt, Brace and Company.
Keynes, J.M. (1936), *The General Theory of Employment, Interest and Money*, Cambridge University Press.
Kopits, G. (2001), "Fiscal Rules: Useful Policy Framework or Unnecessary Ornament?," IMF working paper WP/01/145.
Kopits, G., and S. Symansky (1998), "Fiscal Policy Rules," IMF occasional paper no. 162.
Kremer, M., and S. Jayachandran (2002), "Odious Debt," *Finance and Development*, 39, No. 2, www.imf.org/external/pubs/ft/fandd/2002/06/kremer.htm.
Kuttner, K., and A. Posen (2002), "Fiscal Policy Effectiveness in Japan," *Journal of the Japanese and International Economies*, 16, pp. 536–58.
Laubach, T. (2005), "Fiscal Relations across Levels of Government in the United States," OECD Economics Department working papers no. 462, Paris: Organisation for Economic Co-operation and Development.
Manasse, P. (2007), "Deficit Limits and Fiscal Rules for Dummies," *IMF Staff Papers*, 54, pp. 455–73.
Mankiw, G. (2007), *Macroeconomics*, Worth Publishers.
Masson, P., and M. Mussa (1995), "Long-Term Tendencies in Budget Deficits and Debts," IMF working paper no. 95-128.
Melitz, J., and F. Zumer (2002), "Regional Redistribution and Stabilization by the Center in Canada, France, the UK and the US: A Reassessment and New Tests," *Journal of Public Economics* 86, pp. 263–86.
Mirabeau (1787), *Lettres du Comte de Mirabeau sur l'Administration de M. Necker* (available on the French Bibliothèque Nationale Web site).
Mountford, A., and H. Uhlig (2008), "What are the Effects of Fiscal Policy Shocks?," NBER working papers 14551, Cambridge, Mass: National Bureau of Economic Research.
Mundell, R. (1968), *International Economics*, Macmillan.
Necker, J. (1784), *De l'administration des finances de la France* (available at www.gallica.bnf.fr).
OECD (2002), "OECD Economic Surveys 2001–2002. Japan," *OECD Economic Surveys* no. 18, supp. 2, pp. 1–207.
OECD (2004), "OECD Economic Surveys: United States 2005".
Perotti, R. (1996), "Fiscal Consolidation in Europe: Composition Matters," *American Economic Review*, 86, pp. 105–10.
Pisani-Ferry, J. (2003), "Fiscal Discipline and Policy Coordination in the Euro area: Assessment and Proposals," in *Budgetary Policy in E(M)U, Design and Challenges*, proceedings of a seminar held at the Dutch ministry of Finance.
Reinhart, C., and K. Rogoff (2008), "The Forgotten History of Domestic Debt," NBER working paper 13946, April.
Reinhart, C., and K. Rogoff (2009), *This Time is Different: Eight Centuries of Financial Follies*, Princeton University Press.
Ricardo, D. (1817), *On the Principles of Political Economy and Taxation*, John Murray.
Sachs, J., and X. Sala-i-Martín (1992), "Fiscal Federalism and Optimum Currency Areas: Evidence from Europe and the United States," in Canzoneri M., V. Grilli, and P. Masson (eds.), *Establishing a Central Bank: Issues in Europe and Lessons from the US*, Cambridge University Press, pp. 195–219.
Samuelson, P. (1948), *Economics. An Introductory Analysis*, McGraw Hill (1951).

Sapir, A., Ph. Aghion, G. Bertola, M. Hellwig, J. Pisani-Ferry, J. Viñals, and H. Wallace (2004), *An Agenda for a Growing Europe: The Sapir Report*, Oxford University Press.

Seater, J. (1993), "Ricardian Equivalence," *Journal of Economic Literature*, 31, pp. 142–90.

Spilimbergo, A., S. Symansky, O. Blanchard, and C. Cottarelli (2008), "Fiscal Policy for the Crisis," IMF staff position note, SPN/08/01, Washington D.C.: International Monetary Fund, 29 December.

Stiglitz, G. (2003), *The Roaring Nineties: A New History of the World's Most Prosperous Decade*, WW. Norton & Company.

Sutherland, A. (1997), "Fiscal Crises and Aggregate Demand: Can High Public Debt Reverse the Effects of Fiscal Policy?," *Journal of Public Economics*, 65, pp. 147–62.

Tabellini, G. (1991), "The Politics of Intergenerational Redistribution," *Journal of Political Economy*, 99, pp. 335–57.

Taylor, J. (2000), "Reassessing Discretionary Fiscal Policy", *The Journal of Economic Perspectives*, 14, pp. 21–36.

Taylor, J. (2009), "The Lack of an Empirical Rationale for a Revival of Discretionary Fiscal Policy," *American Economic Review*, 99, 550–55.

US Treasury (2008), "Financial Report of the US Government" (available on US Treasury Web site).

Von Hagen, J., and I. Harden (1994), "National Budget Processes and Fiscal Performance," *European Economies*, Reports and Studies, no. 3.

Western African Economic and Monetary Union (2007), *Multilateral Surveillance Report*, June.

Wyplosz, Ch. (2002), "Fiscal Policy: Institutions versus Rules," CEPR working paper 3238, February.

4

Monetary Policy

4.1 Issues
  4.1.1 What do central banks do?
  4.1.2 The objectives of monetary policy
4.2 Theories
  4.2.1 Principles
  4.2.2 Transmission channels
  4.2.3 Monetary policy in an open economy
  4.2.4 Financial stability
4.3 Policies
  4.3.1 Institutions
  4.3.2 Key policy choices
References

Money is an old device but the concept of monetary policy is relatively recent. Some of the central banks that are in charge of running it are venerable institutions, like the Bank of England which was founded in 1694, but some were only created recently, including the US Federal Reserve, which was founded in 1914. Most central bankers nowadays are very sophisticated policymakers, but their tasks were initially limited to printing and distributing banknotes and coins backed by gold, and to contributing to replenishing the King's coffers. Very few central banks enjoyed real independence in the 1970s, but major reforms occurred in the last two decades of the twentieth century. There has also been considerable advance in the theory of monetary policy. Accordingly, discussions on monetary strategies and policies have evolved a great deal over the last decades.

It is only after the hyperinflation experiences of the 1920s and the subsequent Great Depression that the concept of a macroeconomic role for monetary policy emerged. Indeed, both events have been shown to be related to monetary-policy errors—excessive money creation in the 1920s, excessive money tightening in the 1930s (Friedman and Schwartz, 1971). Those episodes would later lead to a rethinking of the role of monetary policy, but it remained somewhat eclipsed by fiscal policy in the first post-World-War-II decades, a

time when the Federal Reserve was primarily assigned the role of minimizing the cost of public borrowing.

The role of monetary policy was reassessed as a consequence of the mistakes made in response to the inflationary shocks of the 1970s and the subsequent emergence of disinflation as an overriding policy objective. Like the previous episode, this one prompted a deep rethinking of the relationship between monetary policy, growth, and inflation. A lasting consequence of the inflationary mistakes of the 1970s was also that most countries decided to grant independence to their central banks. The way had been opened in 1948 when Germany, remembering the lessons of the hyperinflation episode of the 1920s, created the Deutsche Bundesbank. In most countries, the central bank—once an institution Napoleon wanted to be "in the hand of the government, but not too much"—became a power of its own. By contrast, there was little legal change in the US. Nevertheless, here also the central bank acquired new authority—some would say hubris—thanks to its understanding of financial markets, the design of elaborate strategies, and skillful monetary management.

By the late 1990s, a near-consensus had been achieved that monetary policy had to be mainly geared toward achieving price stability. How this mandate was specified, however, still mattered considerably, and there were subtle differences across central banks as regards the definition of their objectives, their communication and their relationship to government and parliament. Policy discussions therefore were less and less about objectives and more and more about strategies and tactics.

One of the most striking aspects of the evolution of monetary policy has indeed been its increasing sophistication and the growing importance of communication to market participants and private agents. In normal conditions, effectiveness relies heavily on the ability of central bankers to make credible announcements to the public and to steer the expectations of financial-market participants regarding what their future decisions could be. This implies that the impact of monetary policy also depends, sometimes to a considerable extent, on the quality of the central bank's communication.

The financial crisis disrupted in a major way this subtle universe. Starting in the summer of 2007, central banks were immediately propelled to the forefront of the policy response, as they had to react to the crisis of confidence among banks and the drying-up of liquidity in the interbank market (see chapter 8 for a detailed account). To keep the banking system afloat they extended loans to financial institutions in ever-larger quantities and with an increasing risk that they would not be able to recover their money. This highlighted their usually mundane, but nevertheless vital role as guarantors of the smooth functioning of the money market as well as their unique role as lenders of last resort (see below) that are able to step in when private lenders find themselves unwilling or unable to perform their usual function.

As the crisis worsened in the course of 2008, an increasing number of banks found themselves in need of immediate assistance, either because losses incurred on asset markets had made them insolvent, or because market

participants had lost confidence in their financial soundness and would stop lending to them. Central banks temporarily extended emergency lending to distressed financial institutions as a bridge until budgetary support could be provided, and sometimes acted as agents on behalf of treasuries. This highlighted their role as *guarantors of financial stability*.

Finally, the dramatic worsening of the economic situation in autumn 2008 after the bankruptcy of Lehman Brothers, a major US investment bank, led monetary policy to change course. Policy interest rates were sharply lowered, but soon reached the *zero bound** (they could hardly be brought below zero) and several central banks started to engage in *unconventional monetary policy actions*. Beyond short-term lending to banks, these consisted in two main initiatives: First, the direct provision of liquidity to nonfinancial companies through the purchase of short-term securities such as commercial paper. The goal here was to temporarily substitute for a paralyzed banking system. Second, central banks also engaged in *credit easing* or *quantitative easing* and bought longer-term securities such as government bonds in order to keep the asset market operating and lower longer-term interest rates.[1] This illustrated the central banks' mandate to preserve financial stability and their unique power to create money at will to this effect. Unconventional monetary policies began to be gradually unwound when central banks were confident enough that normalisation of economic and monetary conditions was under way.

Central banks are normally proud to be boring institutions, as this highlights their ability to provide stability. The crisis has also indicated that they can on occasions be entertaining ones. This should not lead to overlooking the fact that they also fulfill other, purely technical functions, like the dispatching of banknotes, the supervision of the payment system, or the production of monetary and balance-of-payment statistics.

This chapter starts with a description of what central banks do and a discussion of their objectives. In part 2, we present the modern theory of monetary policy and the lessons that can be drawn from it. The current policy debates are addressed in part 3.

Throughout the chapter we aim to present both how monetary operates in normal times and how it can perform an exceptional role in crisis times. The broader implications of the financial crisis of the late 2000s are addressed in chapter 8.

4.1 Issues

4.1.1 What do central banks do?

a) Liquidity provision

Monetary policy is operated by official institutions called *central banks**, which have the privilege of creating what is called *base money** or sometimes

1. For a presentation of these instruments by central bankers, see Bernanke (2009) and Meier (2009). See also chapter 8.

*high-powered money**. This consists in issuing banknotes and in providing liquidity to the financial system in ways that "maintain price stability and promote a safe and efficient payment system," to quote from the Swedish Rijksbank's fairly standard definition of its tasks. The first task—the issuance of banknotes—is familiar enough, yet of second-order importance in modern economies. Banknotes represent less than 10% of the economically relevant definition of money (see table B4.4.1 in box 4.4). The second task is less familiar, but more important. The best way to understand it is to start from what the central banks actually do on a day-to-day basis.

On any given day, credit institutions (mostly banks) extend credits to households and companies, make payments, and receive deposits from their clients.[2] As these operations do not necessarily balance—some banks are more active in providing credit, others manage a large network of branches where customers hold deposit accounts—banks extend very-short-term loans to each other through what is called the *money market** or the *interbank market**. They are said to provide liquidity to each other. However, the aggregate balance between supply and demand is not left to the market participants alone: The central bank also intervenes on the market by providing its own base money to banks. Also, should they face difficulties in borrowing from other banks, banks can turn to the central bank for the money they need to clear payments, at a fixed price. This ensures both a safe payment system and a stable price of liquidity.

The channel through which this intervention in the money market happens varies from one country to another, but this is immaterial. What is important is that by crediting the account of the corresponding banks at the central bank, the latter provides them with base money which has the privilege of being universally accepted as a means of payment and can be used to settle debts or grant new loans. The central bank supplies enough of this base money to ensure that the financial system runs smoothly and, since it enjoys the privilege of creating base money by the stroke of a pen, it does not face any exogenous limit in the supply of credit.

In practice, liquidity is provided either through *open-market operations**, i.e., purchases of financial assets by the central bank from commercial banks, or through *repurchase agreements** or *repos**, whereby the central bank holds the corresponding assets on its balance sheet for a fixed period.[3] The Federal

2. Financial institutions are regulated and this introduces cross-country differences in their categorisation. Banks in Europe are *universal banks**: the same institutions engage both in retail operations (they collect deposits and extend credits to households and small enterprises), and in corporate finance and merger and acquisition advice. In other words they act both as *commercial banks**, also known as *deposit-taking banks**, and as *investment banks**. In the US, the Glass–Steagall Act of 1933 strictly separated investment banking from commercial banking. In 1999, the Gramm–Bleach–Riley Act authorised the creation of universal banks like Bank of America, Citi or JPMorganChase, but stand-alone investment banks remained until 2008, when they either failed (Lehman Brothers), were absorbed (Merrill Lynch) or registered as deposit-taking banks in order to have full access to Fed refinancing (Goldman Sachs).

3. It is often said that the central bank refinances the commercial bank, hence the notion of *refinancing operation**.

Reserve normally uses the former mechanism whereas the European Central Bank (ECB) uses the latter. In the latter case, the central bank lends new money to commercial banks and receives in exchange financial assets up to exactly the same value that will be recovered if the loan is not refunded (after application of an appropriate discount to the value of the asset—usually called a *haircut**—in order to take into account its quality and protect the central bank from the corresponding market and credit risks). Such assets (known as the loan's *collateral**) traditionally include bills and bonds (both public and private) and in some countries nonmarketable loans and asset-backed securities. Commercial banks commit to buying back these assets after a certain period of time (from one day to a few weeks), hence the name of repurchase agreements.

The designation by the central bank of assets that are eligible as collateral is an important dimension of liquidity management. Before the 2007–09 crisis the range of eligible assets was markedly narrower in the US and the UK (where monetary policy essentially consisted in buying and selling treasury securities on the open market) than in Europe (where the ECB accepted as collateral corporate bonds, loans, and even some high-quality synthetic asset-backed securities). The contrast has narrowed down, however, with the extension by the Federal Reserve and the Bank of England of the range of eligible collaterals. In early 2010, at a time when markets were questioning the viability of Greece's budgetary policy, whether Greek government bonds would remain eligible as collateral for ECB refinancing operations was a matter of life and death for bond-holding Greek banks in desperate need of liquidity. Eventually the ECB decided to keep on accepting them.

The central bank can also influence the banks' lending behavior by asking them to keep a proportion of the deposits received from the public as a deposit with the central bank. This deposit is called a *reserve requirement**. Not all central banks impose reserve requirements, however: Those of the UK, Canada, and Sweden have eliminated them. The ECB does impose a reserve requirement, at a low rate of 2%.[4] In these countries, whether or not banks are required to hold reserves does not significantly affect the conduct of monetary policy. In contrast, the People's Bank of China has been using reserve requirements very actively starting in 2004, raising the reserve-requirement ratio several times a year, as a complement to interest-rates hikes in order to curb money creation in the country.

4. Reserve requirements work the following way. Suppose that a customer has a bank deposit of 100 euros in a bank located within the euro area. Then, the bank must deposit at least 2 euros at the central bank. If the customer uses the 100 euros to repay a debt, then the bank can reduce its reserve accordingly, but the bank of the creditor will raise its reserve deposit by the same amount. In brief, the bank of the depositor can only use 98% of the deposit (here, minus the 2 euros) to extend new loans. In practice, reserve requirements rarely bind quantitatively.

b) The price of liquidity

When drawing liquidity from the central bank, commercial banks pay a fee in the form of a short-term interest rate. For instance, if the rate applied to a repurchase agreement is 3%, a bank seeking liquidity from the central bank for 5 working days for an amount of 100 million euros will have to pay for this liquidity service $5 \times (0.003/360) \times 100$ millions $= 41666.67$ euros.[5] The higher the refinancing rate, the lower the demand for liquidity. Hence by setting a price for its liquidity service, the central bank is able to influence the demand for it. The resulting *money-market rate** will in turn influence all short-term interest rates in the economy and, to a certain extent, long-term interest rates also—and as a consequence the demand for credit and spending and saving behavior.

Although principles are similar, central banks throughout the world do not all operate exactly in the same way to provide liquidity to the banking system, as illustrated by the *modus operandi* of the ECB and the Federal Reserve.

In the euro area, banks normally bid for access to central bank liquidity. The ECB can either allot funds at fixed rate (in which case banks bid weekly for quantities and the ECB sets the interest rate applied to these refinancing operations) as was the case between 2007 and 2010, or it can lend at variable rate. The corresponding rate is normally the minimum rate at which commercial banks can obtain liquidity. This *main refinancing rate* or *refi** is complemented by two marginal financing rates that set a ceiling and a floor to market-rate fluctuations. The three rates are sometimes called *leading interest rates** because they "lead" the market interest rate (see box 4.1).

Every day, the ECB measures an average of interbank rates called the EONIA[6] from a panel of euro area banks. Figure 4.1 confirms that the EONIA fluctuates around the main refinancing rate and that its fluctuations are capped and floored by the two marginal facility rates. This permanent arbitrage mechanism, together with the existence of a unified euro payment system called *TARGET**, ensures the unity of money market rates in the area. Since it is so closely linked to the central bank rate, the call rate is often itself considered a monetary instrument, even though this is not the case.

Box 4.1 The European Central Bank and the Euro Area's Monetary Policy Instruments

The ECB is a federal institution of the European Union whose statute is a Protocol annexed to the EU Treaty. It is managed by an *Executive Board** of six members, including the president and the vice-president. The monetary policy of the ECB is decided by the *Governing Council**, which consists of the Board and the central bank governors of the euro area countries.[a] Implementation is decentralized. It involves both the ECB

5. By tradition the rate on repos is arithmetic, not geometric.
6. EONIA means Euro Overnight Interest Average.

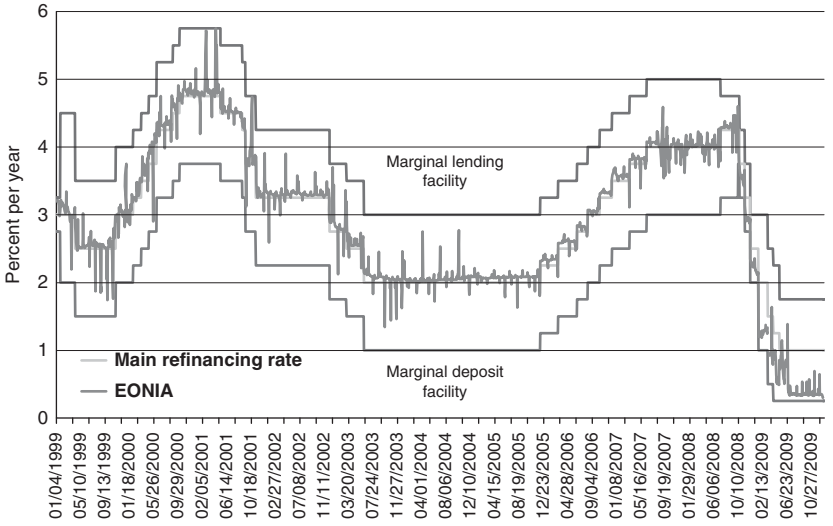


Figure 4.1 Refinancing rates and market rate in the euro area, 1999–2009. Source: European Central Bank.

and the national central banks of the euro area (for example, the Irish banks get refinancing from the Irish central bank and the Dutch ones from the Dutch central bank). The ECB and the national central banks of the euro area together constitute the *Eurosystem**. The *European System of Central Banks* (*ESCB*)* consists of the ECB and all the central banks of the European Union, including those of countries which have not adopted the euro.

The following instruments are used:

- *Minimum reserves* (2% of the demand deposits and of time deposits shorter than two years—including special, regulated accounts). Compulsory minimum reserves are served the main refinancing interest rate.
- Two overnight *standing facilities*: A *marginal lending facility**, in the form of a repurchase agreement at a high rate, and a *marginal deposit facility** remunerated at a low rate. These two facilities ensure that liquidity is always and unconditionally available to banks. A bank seeking short-term liquidity can obtain it weekly through the central bank's main refinancing operations, or at any time at the marginal lending facility rate or by asking another commercial bank (at the overnight *interbank interest rate**, or *call rate**). Similarly, a bank having excess liquidity can deposit it at the central bank at the marginal deposit facility rate or lend it to another bank at the overnight interbank rate. Arbitrage of both types of banks will insure that the overnight interbank rate fluctuates around the main

refinancing rate within a band defined by the two marginal facility rates of the central bank. The overnight interbank rate is a market rate that changes from one transaction to another.
- Weekly *refinancing operations* in the form of competitive bids, through which the Eurosystem provides liquidity to the banks in exchange for public or private securities and loans taken in its balance sheet for two weeks. The corresponding *refinancing rate* is the main rate of the Eurosystem.

In addition, the Eurosystem carries out monthly operations for three-month liquidity for smaller banks and can decide exceptional operations in certain circumstances. On 8 October 2008, in reaction to the worsening of the financial crisis, the ECB decided to serve all bids for liquidity at fixed rate. This full allotment procedure was accompanied by a reduction from 200 to 100 basis points of the width of the standing facilities corridor. As a consequence of this change in the operational framework the *refi* rate became a ceiling for the EONIA rate.

[a] On 1 January 2010 there were 16 countries in the euro area: Austria, Belgium, Cyprus, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Malta, The Netherlands, Portugal, Slovakia, Slovenia, and Spain. All national bank governors, as well as the six Executive Board members, had one vote. The voting mechanism within the Council of Governors is set to evolve as membership grows. See chapter 2.

In the US, the Federal Reserve targets through its open market operations the *federal funds rate** which is the rate at which banks can lend to other banks overnight liquidity from their deposits at the central bank (*Federal funds**). It also sets three discount rates, for primary credit, secondary credit, or seasonal credit, which are available to financial institutions depending on their credit quality (on the principle that the healthiest institutions can get the lowest rate, i.e., the rate on primary credit). The Federal Reserve regularly carries out open-market operations through purchases and sales of US Treasury and securities issued by federal agencies. Finally, there is a reserve requirement of 3% above a certain threshold of deposits, and the percentage is 10% above a second threshold.

Figure 4.2 reports the evolution of the refinancing rates in Germany (or the euro area after 1999), the US and Japan since the 1950s. Three main observations can be made:

1. The refinancing rates have declined over time in line with trend disinflation, and their volatility has also been reduced, especially in the US after Fed Chairman Paul Volcker raised interest rates in 1979–82;
2. There are cycles of rises and reductions, which correspond to phases of economic expansion and slowdown or contraction;
3. In some periods, refinancing rates are kept constant by one or several central banks during several quarters.

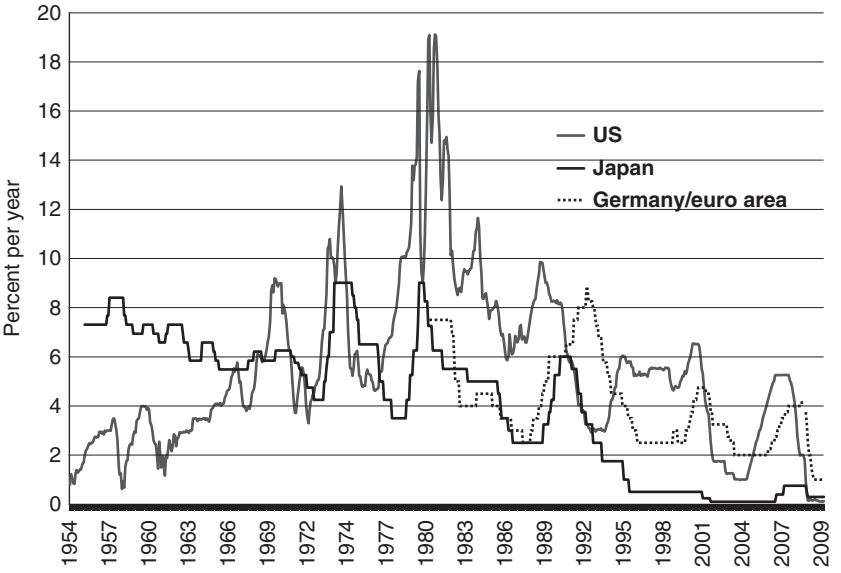


Figure 4.2 Refinancing rates in the US, the euro area, and Japan, 1954–2009.
Source: Central banks.
Note: Germany: Discount rate. Euro area: Rate of the main refinancing operations. Japan: Official discount rate. United States: Federal Funds rate.

The cycles in the three regions were synchronized in the 1970s and the 1980s as central banks reacted to surges in inflation triggered by the oil shocks, but they have become more autonomous in the 1990s and the 2000s (see, for instance, the divergence between the rising US rate and the consistently very low Japanese rate in the late 1990s and 2000s).

c) Liquidity in stress times

Most of the time, banks routinely extend credit to each other and the central bank can limit its role to monitoring this process and to influencing interest rates through the provision to the banking system of limited amounts of liquidity. There are times however when banks are unwilling to lend to each other because potential lenders are uncertain of the ability of the borrowers to repay their debts, or because they themselves prefer to hoard cash in anticipation of future shortages. One such instance was 11 September 2001: Some market participants had had their IT systems disrupted by the attacks on the World Trade Center; others did not know the extent of damage to the IT systems of counterparties; others wanted to keep cash positions at high level in a highly uncertain environment. The Federal Reserve thus feared a liquidity crisis would imperil the economy as a whole. Within hours, it issued a brief statement indicating that "the Federal Reserve System is open and operating. The discount window is available to meet liquidity needs." On 12 September, direct loans to private banks amounted to $45bn (against $0.19bn on the

same day of the previous week), and in the following days the Federal Reserve flooded the market with liquidity through buying record amounts of securities in open-market operations.

Severe financial shocks also give rise to liquidity crises. When the extent of the US subprime credit crisis began to be realized in summer 2007, the fear that major banks would face funding problems or even bankruptcy as a consequence of the depreciation of financial products held in their portfolios started to spread among market participants. As the losses had not been disclosed yet, each bank started to value counterparty risk and the market for interbank liquidity provision came to a standstill (box 4.2).

Box 4.2 The Onset of a Financial crisis: Liquidity Stress in 2007 and 2008

On 9 August 2007, French bank BNP–Paribas announced that it could not fairly value the underlying assets in three funds open to retail investors as a result of their exposure to US subprime mortgage lending markets. This announcement triggered fears about the financial situation of major interbank market participants and a significant deterioration of liquidity in the US and Europe

How could a problem on the real estate market affect the interbank market? Two mechanisms were involved:

- First, many banks faced the risk of having to provide funding to specialized investment vehicles they had created and guaranteed (see chapter 8).
- Second, as an increasing number of banks reported losses or potential losses, banks became increasingly unwilling to provide funds to counterparties in the interbank market, where banks lend short-term to each other without requiring the posting of collateral (lending is *unsecured**). A climate of distrust and uncertainty developed and this in turn caused a spiraling of the banks' perceptions of *counterparty risk** (associated with the default of the borrower) and *liquidity risk** (the risk of not having access to liquidity or having to pay an excessive price for it). The hoarding of liquidity led to a marked weakening of interbank money market activity.

Liquidity conditions are often measured by spreads between the interbank market rate and the rate on government securities of identical maturity (the so-called *TED spread**), or by the spread between the three-month interbank market rate and the capitalized overnight swap rate (OIS), which measures the premium over the markets' expectations of future policy rates (the so-called *OIS spread**). Neither measure is perfect but both aim at capturing the tension on the interbank market.

While these spreads had until summer 2007 been inferior to 50 basis points in the US and close to zero in the euro area, they edged upward in

early August 2007 and remained for several months above 100 basis points in the US and above 50 basis points in the euro area. In September 2008, after the collapse of Lehman Brothers, they rose even further, at levels indicating a near-total paralysis of the interbank market and gradually abated only after governments had announced bank rescue packages (see figure B4.2.1).

In the euro area, a further indication of the dislocation of the interbank market and the resulting collapse of liquidity during the financial crisis were the rise in the dispersion across countries of interbank rates such as the *Euribor* rate* (the rate at which banks offer to lend unsecured funds to other banks). Whereas its standard deviation is normally about 1 basis point (0.01%), it rose to more than 15 basis points in November 2008. Even within-country standard deviation exceeded 10 basis points. This did not mainly result from the pricing of counterparty risk as the standard deviation of collateralized loans also rose very significantly.

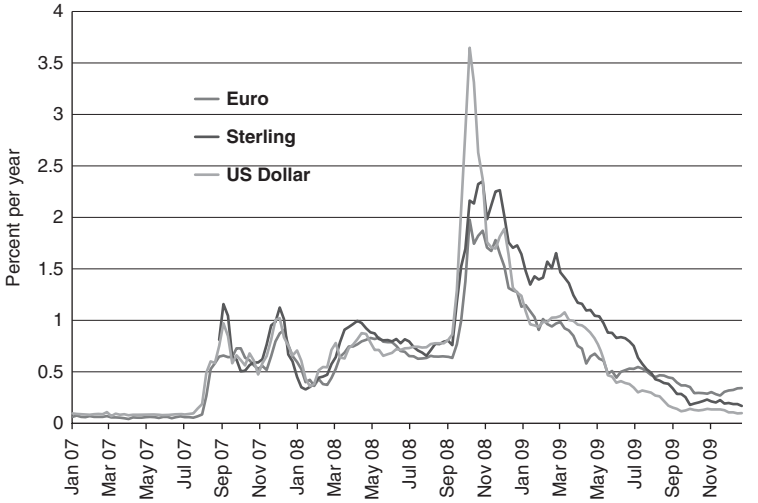


Figure B4.2.1 OIS spreads in the euro area, the US, and the UK.
Source: Reuters.

Such episodes help understand what is meant by *liquidity**. An important distinction is to be drawn between:

- *Market liquidity**, which can be defined as the ease with which a position in an asset can be liquidated without appreciably altering its price. Threats to it arise when assets that are normally traded in reasonable sizes with little price impact can only be transacted at a substantial premium or discount, if at all. The concept is *asset-specific*.

- *Funding liquidity**, which can be defined as the ease with which a solvent institution can service its liabilities as they fall due. Illiquidity occurs when solvent counterparties have difficulty borrowing immediate means of payment to meet liabilities falling due. This concept is *institution-specific*.

The two types of illiquidity are distinct but interdependent because illiquidity of a given market affects institutions which are heavily involved in it, and vice-versa. The crisis in 2007 started as a market liquidity crisis affecting mortgage-related assets and quickly became a crisis of the funding of institutions with significant exposure to the mortgage market.

Central banks assume a crucial role vis-à-vis both categories of risks, especially when funding strains imperil the viability of financial institutions.

In 2007–08 it quickly became apparent that banks were not willing to use the discount window, as requesting access to it would have involved the risk of signaling to the market a state of financial stress—and thereby of worsening further its access to credit.[7] Rather, central banks engaged in massive direct financing operations. In a first stage, they were able to increase the volume of long-term refinancing to the market without expanding their balance sheets by withdrawing liquidity at other maturities. This was intended to preserve monetary policy from being affected by the provision of emergency liquidity. In autumn 2008, however, both central banks renounced trying to limit the size of their balance sheet and embarked on outright credit expansion (see chapter 8).

d) From short-term to long-term interest rates

Due to banks' arbitrage, short-term market interest rates always remain close to official rates. They also influence interest rates for longer maturities, albeit in a far from mechanical way.

The *yield curve** (i.e., the interest rate as a function of maturity) is primarily affected by expected monetary policy. This is because portfolio managers who want to invest over a long period can either hold long-dated assets or roll short-dated assets over time. If they are not averse to risk, the long-dated interest rate should be the average of the sequence of expected future short-dated interest rates (box 4.3). Suppose investors expect short-run interest rates to increase in the future. In this case, they will temporarily prefer buying short-run assets in order to benefit from the future interest-rate rise. This will push long-run interest rates upward compared to short-run ones, and the yield curve will be steeper. In the reverse case (expected interest-rate fall), the yield curve will be flatter or even downward sloping (*inverted yield curve**). On 5 April 2010 (figure 4.3), the yield curve was steeply upward-sloping in major advanced economies as markets expected gradual exit from near-zero

7. There were several instances when UK banks were reported by the press to have borrowed from the Bank of England's discount window. This created uncertainty as the central bank would not disclose the names, amounts, and reasons.

policy rates. Expectations of future rates were lower in Japan and Germany than in the US and the UK, resulting in lower long-term rates.

Real-world investors are risk-averse:[8] Investments with a longer maturity have a more uncertain return, hence the existence of a risk premium called the *term premium** embedded in longer-term interest rates. Consistently, even when no change in short-term interest rates is expected, the yield curve is generally upward sloping: Short-run interest rates are those targeted by the central bank, and longer-term rates are higher. Inverted yield curves are exceptional events that can be observed only when a sharp fall in the interest rate is expected (for example, as a result of successful monetary contractions).

Box 4.3 The Yield Curve

Most bonds pay a fixed interest rate and are therefore called *fixed-income securities**. They provide a regular (typically, annual or semi-annual) payment called a *coupon**, and the coupon rate is the ratio of this coupon to the borrowed amount, or *principal**, which is to be refunded at maturity. When issued, bonds are traded on financial markets and the market interest rate is defined as the internal rate of return of the bond given its market price. There are a whole range of possible maturities, and hence of interest rates, from a few weeks to 50 years. The standard theory of the yield curve relies on investors arbitraging between a long-term investment (paying the long-term rate) and a succession of short-term investments (each one paying the corresponding short-term interest rate). As the long investment is riskier (holding the bond until it expires involves an inflation risk, liquidation before the term involves a *capital risk**[9]), the long investment generally yields higher interest than the succession of short investments. More precisely, the interest rate for maturity N, i_t^N, can be expressed as a function of expected short-term rates $i_{t+\tau}^1$ and of a term premium ρ_t^N. Thus:

$$(1 + i_t^N)^N = (1 + i_t^1)(1 + E_t i_{t+1}^1) \dots (1 + E_t i_{t+N-1}^1)(1 + \rho_t^N)^N \quad \text{(B4.2.1)}$$

where $i_{t+\tau}^1$ is the one-year interest rate in $t + \tau$ and ρ_t^N is the annualized term premium, defined as the extra return that is required by investors to compensate for holding riskier assets. The term premium grows with N. Hence, the yield curve is generally upward sloping—absent expected interest-rate variations. It is important to note that the expected interest rates are not directly observable; therefore the term premium is not observable either. However, future interest rates are traded on forward markets and this allows it to be evaluated.

8. Risk aversion is defined in chapter 2.
9. i.e., the risk of a fall in the market price of the bond before its liquidation.

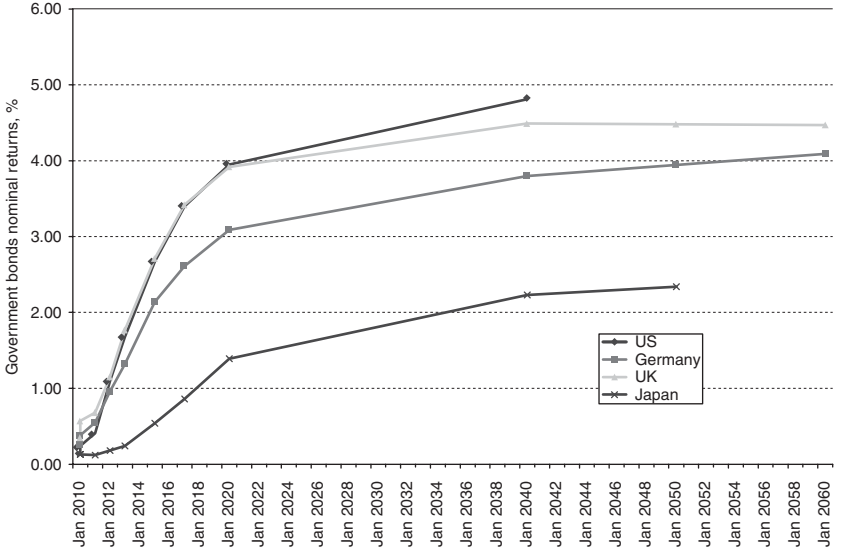


Figure 4.3 The yield curve on 5 April 2010.
Source: Reuters.
Note: The yield curve is based on government bond nominal returns in the countries

Because longer-run interest rates incorporate expectations concerning future monetary policy, they can change even when short-run rates are held constant. Central banks nowadays tend to avoid creating surprises and they use this property to smooth the evolution of long-run interest rates by communicating their intentions through speeches and interviews. For instance, figure 4.4 shows that the successive hikes of the main refinancing rate by the ECB in 2006 were incorporated in interest rates of one-month maturity or more before they took place. Indeed, longer-term interest rates rose smoothly through the year.

e) Nominal and real interest rates

A familiar and important distinction exists between nominal and real interest rates. For each maturity, the *real interest rate** is the difference between the *nominal interest rate** and the expected inflation rate over the same period. Because the expected—rather than observed—inflation rate enters into its determination, it is sometimes called the *ex ante* real interest rate, while the difference between the interest rate and observed inflation is called the *ex post* real interest rate. Both notions can be used but only the *ex ante* real rate matters for economic decisions.

The (*ex ante*) real interest rate can be estimated using surveys and forecasts of future inflation, or it can be deducted from prices on financial markets. Some governments issue inflation-protected bonds whose principal and coupons are indexed on the consumer price index and which therefore deliver

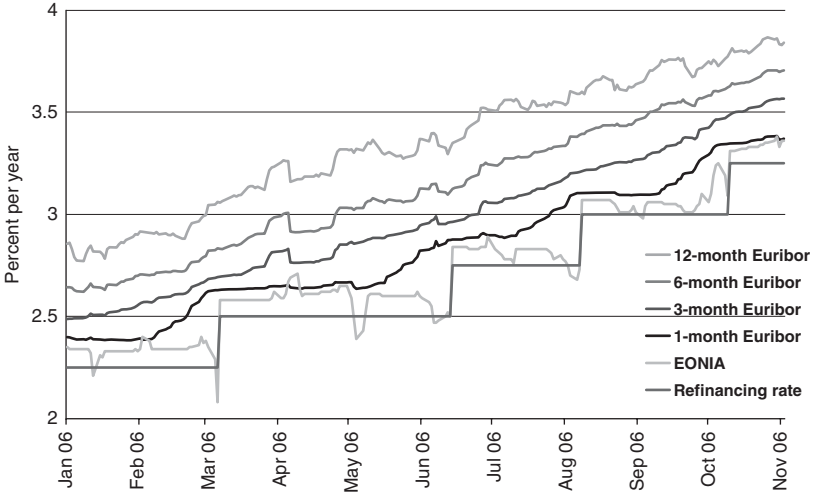


Figure 4.4 Market and policy interest rates: The euro area in 2006.

a real yield to their holder (*Treasury Inflation-Protected Securities** or *TIPS* in the US, inflation-linked gilts in the UK, OATi and OAT€i in France, etc.). The comparison between the return on those bonds and the return on conventional, nonindexed bonds of equal maturity is frequently used to gauge the inflation expected by market participants. For example, in March 2007 in the euro area, the yield of a conventional government bond maturing in 2015 was 3.96% while the yield of an inflation-protected bond with the same tenor was 1.84%, suggesting an inflation expectation of $3.96 - 1.84 = 2.12\%$ per annum over the period 2007–15.[10] However, this measure is blurred by the limited liquidity of the market for inflation-protected securities (Hördahl, 2009).

f) International linkages

Capital mobility across countries blurs the link between monetary policy and interest rates. This is because investors can arbitrate not only between short-run and long-run assets, but also between domestic and foreign assets. For instance, the long-term rates in the euro area and in the US depend on expectations concerning future monetary policy. However, for each maturity, investors can arbitrate between euro area and US assets. This makes the interest rates across the Atlantic interdependent (figure 4.5).

10. More precisely, the difference between the two yields, otherwise known as the *break-even inflation**, is the sum of inflation expectations and a term premium specifically linked to inflation, the *inflation premium**. The inflation premium is difficult to measure but it is believed to be relatively stable over time. Movements in break-even inflation can thus be interpreted as changes in inflation expectations.

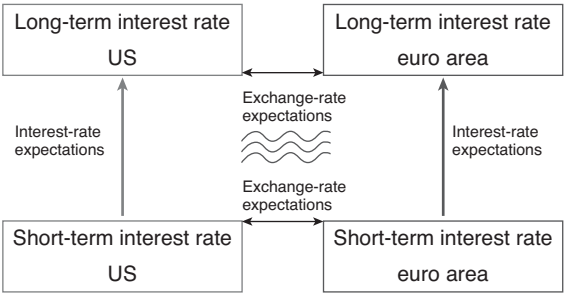


Figure 4.5 International interest rate linkages: A stylized view.

Interdependence does not mean identity, for two reasons. First, some countries are considered riskier than others because of higher indebtedness, political uncertainty or weak legal protection and financial supervision. Hence a *country-risk premium** is added to their interest rates in comparison to less risky countries, especially for long-run assets. Country-risk premiums can reach several percentage points in emerging countries and in countries experiencing financial stress. In the spring of 2010, the interest rate differential between Greek and German government bonds, both denominated in euro, exceeded 5% per annum as fears of Greek default rose. Second, interest rates differ across countries if the exchange rate is expected to vary. This is because investors will require a higher return from an asset denominated in a currency that they expect will depreciate. For instance, for any given asset class, the interest rate will be higher in the US than in the euro area if the dollar is expected to depreciate against the euro. We shall come back to this relationship, called uncovered interest-rate parity, which is very important for the conduct of monetary and exchange-rate policy.

g) What about money?

So far, we have refrained from mentioning the quantity of money in circulation. However, it has played an important role in theory and policy debates and some central banks maintain objectives for growth in monetary aggregates.

Money is hard to define and even harder to measure (box 4.4). The concept is simple—*fiat money** consists in a deposit at a bank (or a similar institution) that can be used together with notes and coins as a medium of exchange—but as financial innovation has developed, there is now a continuum of financial instruments which meets this definition.

Historically, *monetary aggregates** corresponding to various definitions of money have played an important role in the discussion about monetary policy. In the 1980s, most central banks relied on such aggregates to guide policy. They were essentially used as observable intermediate objectives that were supposed to be strongly correlated with future inflation as the quantity theory

of money would predict (see section 4.2). A central bank targeting low inflation would thus define a path for monetary aggregates consistent with its price-stability objective. Money would thus serve as a leading indicator of future inflation. However, experiences with strict control of monetary aggregates, especially in the US and the UK in the late 1970s, resulted in high interest-rate instability, and monetary aggregates proved to be poor predictors of inflation in a financial-innovation context. Aggregates were thus put aside as policy indicators and the US Federal Reserve has even stopped publishing some of them. Nevertheless, the discussion has not ended. The European Central Bank remains more faithful to the aggregates than the Federal Reserve or the Bank of England. We shall return to this discussion in section 4.3.

Box 4.4 Money and Monetary Aggregates

The traditional definition distinguishes between the money directly circulated by the central bank (coins and notes in circulation plus deposits of commercial banks at the central bank), which is called $M0^*$ and is registered as a liability of the central bank, and money issued by commercial banks for their customers. However, while it is clear that a deposit on a cash account is being used for the purchase of goods and services and is therefore equivalent to bank notes, should a savings deposit that can be transferred overnight into the cash account also be regarded as money?

Various monetary aggregates have thus been defined: $M1^*$ includes both M0 and demand deposits. Hence M1 is the sum of the most liquid liabilities of both the central bank and commercial banks. Similarly, $M2^*$ includes M1 and deposits with a maturity of up to two years, whereas $M3^*$ is the sum of M2 and of money market instruments, i.e., marketable securities with less than one year to maturity (table B4.4.1).

Table B4.4.1
The money aggregates of the euro area, in billions of euros and in % of M3 in February 2010

M1	Currency in circulation	
	Overnight deposits	
	Total	4565 (49%)
M2	Deposits with an agreed maturity of up to two years	
	Deposits redeemable at notice of up to three months	
	Total	8225 (88%)
M3	Repurchase agreements	
	Money market fund shares/units	
	Total	9321 (100.0%)

Source: ECB, *Monthly Bulletin*, April 2010

The central bank creates money at will (table B4.4.2). This happens when it provides liquidity to a commercial bank through buying a financial asset (for example, a government bill) or receiving it in a repurchase agreement: The assets-and-liabilities side of its balance sheet increases by the corresponding amount, say, 100. The commercial bank, in turn, replaces the government bill by central bank money on the asset side of its balance sheet.

Table B4.4.2
Money creation by the central bank

Central bank		Commercial bank	
Bills 100	Money 100 ↔	Bills: 100 Money: 100	Deposits 100

Note: Assets are on the left, liabilities on the right. Total money created: 100 by the central bank.

Commercial banks also create money. For example, a commercial bank extends a credit of 100 to a customer, who in turns spends it on goods and services. This implies that the customer draws on his deposit account for, say, 80, and transfers the corresponding money to the accounts of other customers in other banks. The bank which initially extended the credit retains at that point only a fraction of the initial deposit (in this example, 20). The other banks receive the deposits of the other customers (80), which can be used to extend new loans (table B4.4.3). There is money creation each time the banking sector extends a loan to nonbank customers, because this amounts to increasing the total amount of deposits in the system.

Table B4.4.3
Money creation by commercial banks

Central bank		Commercial bank		Customers	
Claims on commercial banks 100	Money 100	Loan: 100	Customer deposit: 100 Customer deposit: 80 20	Bank accounts: 100 Bank accounts: 80 20	Debt: 100

Note: Assets are on the left, liabilities on the right. Total money created: 200, of which 100 by the central bank and 100 by commercial banks.

If commercial banks extend loans in constant proportion to money received from the central bank, the ratios of M1, M2, and M3 to M0 are constant and called *money multipliers**. Control of M0 thus allows the central bank to control the total amount of money in circulation. However, the link between M0 and other aggregates has considerably loosened over time, especially because close substitutes to the least liquid components M2 and M3 have emerged as a consequence of financial innovations.

4.1.2 The objectives of monetary policy

The objectives that central banks should pursue constitute their *mandate**. These have varied significantly over time and are still a matter for discussion among politicians and economists. In the 1970s, it was common for central banks to have broad mandates involving difficult trade-offs between alternative targets. One of the lessons drawn from the inflation of the 1970s and the 1980s has been that central banks ought to be given more precise objectives; price stability emerged as the dominant one. However, not all central banks have a mandate focused on price stability and even those that do may have to pursue other objectives simultaneously. In addition, the financial crisis of 2007–09 has opened a discussion on whether central banks should be less focused on controlling price inflation and gear monetary policy more towards financial stability.

a) Price stability

Pursuing price stability amounts to maintaining the real value of money, that is, its *purchasing power**: The quantity of goods, services or assets that one unit of money can buy. More precisely, it amounts to maintaining its internal value (its purchasing power in terms of the domestic consumption basket), which has to be distinguished from its external value (the purchasing power in terms of foreign currencies).

The justification for assigning a price-stability objective to the central bank is threefold: First, price stability is a desirable objective from a social welfare point of view (the "*what*" question); second, central banks are best placed to reach this objective (the "*who*" question); third, assigning any other task to them would distract them from accomplishing the former.

The benefits of price stability are rather intuitive although, as noted by Buiter (2006), their derivation from theory is not straightforward. The most frequently mentioned is that inflation distorts economic decisions through the implicit taxation of cash balances and the blurring of relative price signals. This is why most central banks aim at keeping the *inflation rate**, i.e., the annual increase in the general level of prices, at a low value. What exactly this low value should be is a delicate question to which we shall return in section 4.3.

The answer to the *who* question is not obvious either. The *monetarist** answer is best captured by Milton Friedman's famous sentence "inflation is always and everywhere a monetary phenomenon" (Friedman and Schwartz, 1971), which points to a direct causal relationship between the quantity of money in circulation and inflation. This proposition implies that price stability requires controlling the amount of money in circulation and makes monetary policy the natural instrument for controlling inflation. However, as we shall see in section 4.3, the medium-term direct relationship between money and prices has broken down in recent times, and contemporary economic models of the kind we will present in section 4.2 do not give a special role to money. There must therefore be other justifications for assigning the control of inflation to monetary rather than, say, to fiscal policy.

The arguments are both economic and institutional. First, contemporary economic models retain an important assumption called the *long-term neutrality of money**, i.e., the disconnection, in the long run, between nominal variables (such as the general level of prices, nominal wages, interest rates, the nominal exchange rate . . .) and real variables (real GDP, employment, real wages, real interest rates, the real exchange rate . . .). Though it has real effects in the short run, over a long horizon, monetary policy can best control nominal variables without affecting real variables. This is not the case for fiscal policy, which affects the composition of output both in the short run and in the long run. Second, controlling inflation should not be distracted by other policy objectives that may influence the price level, such as output targeting or the financing of public deficits (except insofar as they help to predict inflation). Independent institutions with a narrow mandate are better equipped to do this. For those reasons, the central bank has been put in charge of price stability in each and every country.

Central banks have been spectacularly successful in reaching this objective. Figure 4.6 shows the distribution of inflation rates in the world from 1980 to 2008. In 1980, less than 10% of countries had an inflation rate lower than 5%. This proportion produced 60% in the early 2000s and in 2008, in spite of a worldwide inflation push, it was still about 30% and only around 15% of all countries experienced inflation above 15% a year.

Indeed, the near-disappearance of inflation was a characteristic of the 1990s and the early 2000s, and even the dramatic increase in oil and raw material prices in the mid-2000s did not provoke a major inflationary fever as had been the case in the 1970s.

In the early 1980s, central banks inherited high inflation rates as a result of the two oil shocks experienced in the 1970s, which had been amplified by wage indexation as well as expansionary policies. As a consequence, a number of central banks tightened monetary policy either through discretionary policies (e.g., the Federal Reserve) or through anchoring to low-inflation countries (as many European countries did vis-à-vis Germany). Still, some emerging countries registered very high inflation rates in the 1980s. In Israel, for instance, annual inflation exceeded 100% from 1980 to 1986. Some countries

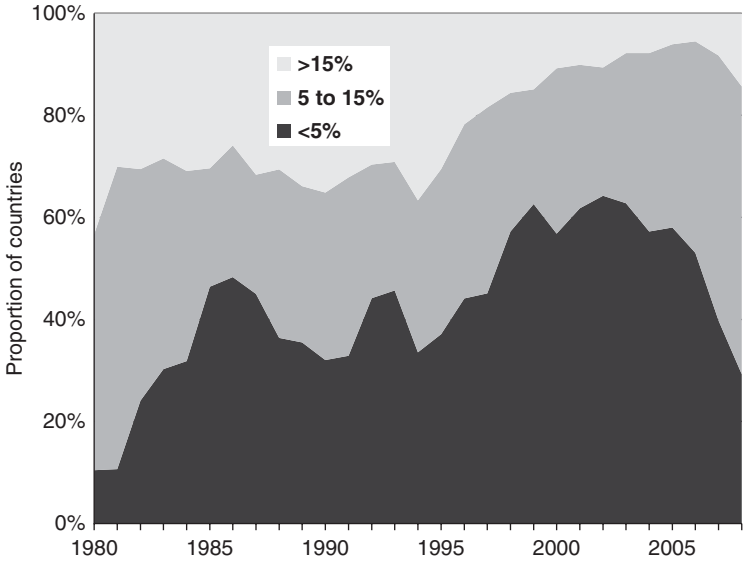


Figure 4.6 World distribution of inflation rates, 1980–2008.
Source: International Monetary Fund.

even experienced *hyper-inflation**, which is usually defined as an inflation rate higher than 50% per month (Cagan, 1956).[11] For instance, prices increased by 20266% in Argentina between March 1989 and March 1990.

A major achievement of the 1990s was disinflation—though Japan overdid it and experienced *deflation**, that is, a joint fall in output and the price level. This phenomenon had been observed in the interwar period but was considered a historical curiosity. The Bank of Japan was initially slow to react, until it set interest rates to zero and started to aggressively create money, eventually engineering growth and inflation.

How much of the price stability observed in the 1990s and the early 2000s was due to favorable worldwide conditions and how much to the quality of monetary policies and institutions is hard to tell. In fact, two major explanations are given for the overall reduction of inflation of the recent decades. One is the generalization around the world of the central bank independence model—an institutional development. At least as important, albeit more recent, is globalization and the release on the global market of a huge production capacity from China and other developing countries—a positive worldwide supply shock. The resurgence of inflation as a consequence of the large increase in oil and food prices experienced in the mid-2000s—a worldwide negative supply shock—is an indication that institutions cannot take credit for the whole disinflation performance. The fact that the rise in

11. 50% per month corresponds to 12875% a year.

consumer prices remained limited in spite of the magnitude of the shock is an indication that they nevertheless deserve some credit for it.

b) Exchange-rate stability

An historically important role of monetary policy has been exchange-rate stability. Until the 1990s, many countries relied on a fixed exchange rate as a means of controlling inflation and, after the demise of the Soviet bloc, several countries in transition decided to "anchor" their economy through the setting of a fixed exchange rate. As recently as in 1995, in many countries in Europe—France, Spain, Belgium, and The Netherlands, for example—or elsewhere—Argentina and Brazil—monetary policy was entirely geared toward maintaining the external value of the currency vis-à-vis some larger country: Germany in Europe, the US in Latin America. The attraction of fixed exchange-rates has faded away in recent years: Apart from China, only smaller European countries such as Denmark, some Carribean countries, and former French African colonies continue to peg their exchange rates. These countries chose (or still choose) to stabilize the external rather than the internal value of the currency. For small and open countries, the two objectives (internal and external value) are closely related since imported items weigh heavily in the domestic price index.

Other countries, notably in Asia, do not formally target the external value of the currency but nevertheless attempt to limit exchange-rate fluctuations. This represents a constraint on the ability of monetary policy to maintain price stability and involves a trade-off between the internal objective and the exchange-rate objective. This tension was especially visible in the case of China in the 2000s. We shall return to the choice of an exchange-rate regime in chapter 5.

c) Output stabilization

Like fiscal policy, monetary policy has a short-run impact on aggregate demand. This is because in the presence of price rigidities a lower interest rate tends to encourage investment (through a lower real interest rate) and net exports (through a depreciated real exchange rate), and because higher prices reduce the purchasing power of those assets, like conventional fixed-rate bonds, that are not perfectly indexed to inflation. Monetary policy can therefore be used to stabilize aggregate demand, i.e., support demand through an *expansionary monetary policy** when demand is weak and a *restrictive monetary policy** when demand is ballooning.

The rationale for such *counter-cyclical** monetary policy goes back to the Great Depression of the 1930s but, as for fiscal policy, the desirability and the effectiveness of counter-cyclical monetary policy are debated. As will be

detailed in section 4.2, the existence of price rigidities, a hypothesis upon which counter-cyclical monetary policy relies, is not much debated anymore. However, the long and variable lags involved in the transmission of monetary-policy impulses make discretionary stabilization a delicate exercise and may transform a counter-cyclical policy into a procyclical one. This is why the degree of central bank activism is a matter for discussion. Market expectations may also impede counter-cyclical policy through the adjustment of long-run interest rates. For example, the long-term interest rate may increase in a recession if short-term rates are lowered very aggressively and are expected to lead to future inflation.

Central banks behave in practice as if they were aiming at minimizing the output gap. In 1993, John Taylor showed that the average reaction of the Federal Reserve to US inflation and the output gap could be captured by the following simple equation:

$$i_t = \bar{r} + \pi_t + 0{,}5(\pi_t - \tilde{\pi}) + 0{,}5(y_t - \bar{y}_t) \qquad (4.1)$$

where i_t is the short-term, nominal interest rate, π_t the inflation rate, $\tilde{\pi}$ the inflation objective, $(y_t - \bar{y}_t)$ the output gap (difference between output and its potential level, see chapter 1), and $\bar{r}$ the "neutral" level of the real interest rate.[12] Such behavior was later confirmed for other central banks (see Bernanke and Mihov, 1997, for Germany). Equation (4.1), called the *Taylor rule**, has become one of the economists' basic tools to assess interest-rate variations.

Although it has no normative content, the Taylor rule is a useful standard for comparing monetary stances over time and across countries. For instance, figure 4.7 compares the evolution of the short-term interest rate of the euro area with a Taylor rule. According to this graph, the ECB was rather accommodating for most of the period, compared to a Taylor rule. From 2005 to mid-2008, however, there was a divergence between the Taylor rule based on headline inflation and that based on *core inflation** (i.e., consumer price inflation excluding fresh food and energy). This divergence was due to the combination of a large increase in energy prices and the absence of any "second-round" effect on output prices and wages. It created a policy dilemma for the ECB since it was difficult to determine whether core inflation would converge upward or headline inflation downward. By raising interest rates in 2005, at a time when this was not called for by the headline inflation rule and by raising them moderately only in early 2008, at a time when headline inflation was ratcheting upward, the ECB has adopted a middle road. It has in fact explicitly indicated that its policy was not to respond to headline inflation but to prevent second-round effects.

12. The "neutral" interest rate can be defined as equal to the growth rate of the economy, which maximizes consumption per capita at the steady state according to the golden rule of growth theory (see chapter 6).

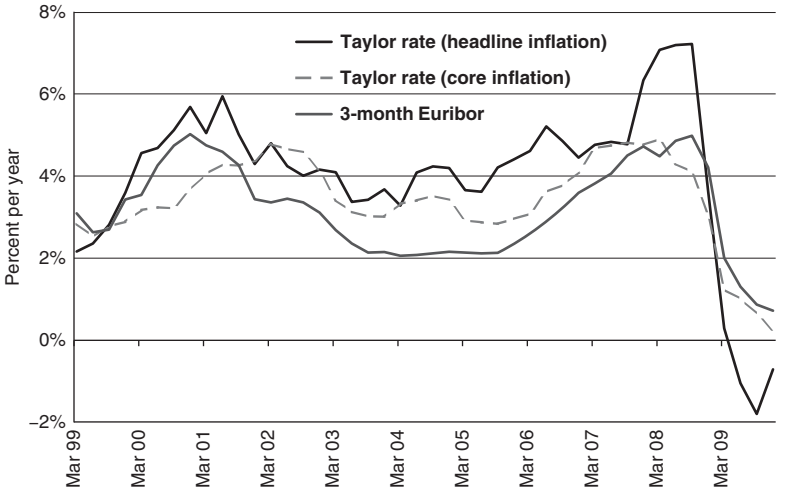


Figure 4.7 The Taylor rule in the euro area, 1999–2009.
Source: Authors' calculation based on OECD data.

The fact that central banks appear to react to the output gap does not imply that they have an output-stabilization objective. As a measure of excess supply of goods and services in the economy, the output gap is a predictor of future inflation. Raising the interest rate is the appropriate reaction to curb future inflation when demand exceeds potential output, even for a central bank that does not pursue output stabilization per se.

d) Financial stability

Financial stability, i.e., the proper functioning of banks and financial markets, was not a major concern in the context of the highly segmented and regulated post-Great-Depression financial systems of the 1960s but after liberalization unleashed market forces again in the 1980s and the 1990s, the issue gained prominence again.

Responsibility for financial stability is generally shared between regulatory agencies that deal with one or several specific market segments (such as securities, banking, insurance, etc.), the central bank, and the Treasury. Arrangements vary from country to country and over time as no best model has yet emerged, in particular as regards the role the central bank should play. The responsibility of regulators and supervisors (in charge of enforcing regulation and of overseeing individual banks) is microeconomic in nature whereas the central bank's is macroeconomic (the modern jargon opposes micro-financial and macro-financial aspects to emphasize the specific financial dimensions of the issues at hand). A proper micro-financial framework involves inter alia the setting of standards in order to ensure that

banks properly manage the risks they are taking and hold sufficient capital to cover them. This is the role of *prudential policy**. It is a necessary condition for financial stability but it is by no means a sufficient one: Even sound financial system are subject to bubbles.

The oversight of banks was traditionally the cornerstone of financial stability. However, the scope for it widened in the 1990s with the expansion of securitization (the transformation of loans into marketable securities that can be sold by banks to other market participants, see chapter 8) and the growth of *financial derivatives** (financial instruments whose value is determined, often in a nonlinear way, by the evolution of the price of a given asset).

Because it acts through changing the relative price of present and future consumption as well as the incentives to invest, monetary policy heavily relies on the banking and financial sectors that pass monetary impulses onto credit and market interest rates. Therefore, a safe banking and financial sector is crucial for monetary policy transmission and central banks are very much concerned by financial stability. This can lead them to extend large amounts of liquidity to the banks in the short run when all of them are simultaneously seeking liquidity, and therefore cannot lend to each other.

The reason why central banks are willing to provide liquidity to markets in times of stress is that events that endanger the ability of some borrowers to meet their obligations may degenerate into a chain reaction—what is called a *systemic crisis**. On 9/11, the fear was that some market participants would simply not be able to participate in transactions. Similar dangers arise from the default of a large or very interconnected borrower whose default puts in danger the solvency of institutions heavily exposed to it. This was dramatically illustrated by the consequences of the default on 15 September 2008 of Lehman Brothers, the investment bank, after the US government reversed its previous stance and decided not to bail it out on moral hazard grounds (box 4.5).

Box 4.5 The Consequences of the Lehman Bankruptcy[a]

Lehman Brothers, the US investment bank, filed for bankruptcy on Monday, 15 September 2008. It had suffered significant losses resulting from its exposure to troubled subprime-related assets and the market had come to the assessment that its efforts to raise capital had not been sufficient to cover declared and future expected losses. In the days prior to bankruptcy the spread on its *Credit Default Swap** (or CDS, the cost of insuring against its default) had reached 600 basis points, indicating a high perceived probability of default. Its access to liquidity was consequently cut off. During the weekend of 13–14 September, government-sponsored discussions about a possible take-over had not succeeded and the US Treasury had refused to engage in government bail-out, making bankruptcy unavoidable.

At the time of collapse Lehman's consolidated debt amounted to more than $600bn (more than 4% of US GDP) and it held a central position as a dealer and counterparty in a variety of compartments of the financial market. In the following days its default had major repercussions:

- On the CDS market the default clause of all contracts referencing Lehman was activated (meaning that the provider of insurance had to pay its counterparty) and all contracts in which Lehman was a counterparty (as a buyer or a seller of insurance on, say, a possible default of Goldman Sachs) were immediately terminated. However, as these contracts were essentially of the over-the-counter type, no public information on their volume was available, which created major uncertainty about the size of the shock and its implications for individual financial institutions and markets. Also, there was no netting of positions (although a special trading session had been organized on the Sunday to allow major dealers to net out their positions). It later appeared that the nominal amount outstanding of contracts referencing Lehman was $72bn and that corresponding net exposure amounted to a modest $6bn, but in the meantime the default had had major implications on the CDS market volatility. Furthermore, coming a few months after the demise of Bear Sterns, Lehman's failure was regarded as a signal that the business model of investment bank was vulnerable and that the US government was ambiguous about bail-outs. The results were a major rise in the CDS spreads of investment banks.
- The shock reverberated on the money market. Lehman was a major issuer of short-dated debt and its paper was considered attractive by funds investing in money markets. In the aftermath of the Lehman bankruptcy investors shunned commercial paper and other forms of short-term debt, prompting Fed action to substitute private investors with purchases of short-term private debt.
- Lehman was also a major broker–dealer of securities. As a consequence of the bankruptcy procedure, investors that had placed investment assets with Lehman's broker–dealer units to serve as collateral lost access to these assets (at least for the duration of the procedure). This prompted the liquidation of other assets.

This immediate, quasi-mechanical impact resulted from the fact that Lehman, though not especially big, was very interconnected. The same would have applied, but with a different order of magnitude, to American International Group, an insurer that was bailed out by the US government a few days after the Lehman failure. Beyond the mechanical impact,

the refusal by the US government to bail-out Lehman had the broader consequence of signaling that bankruptcy of a well-known financial player was a possibility. This resulted in an across-the-board repricing of risk.

[a]This box is based on the Bank of International Settlements (Fender et al., 2008).

The financial stability role of the central banks raises three policy issues which are a matter of ongoing discussion:

- *Moral hazard.* Through acting as a *lender of last resort** that extends assistance to systematically important financial institutions when they find themselves unable to raise money on the market, the central banks (and the treasuries to the extent they follow suit and bail-out insolvent institutions) may encourage imprudent behavior. Furthermore, the collateral provided by illiquid financial institutions in the context of repurchase agreements may be of inferior quality, which may imply that the central bank de facto engages in implicit bail-out.[13] This is a classic dilemma in the theory of insurance (already mentioned in chapter 2) that is fully relevant for the case of emergency liquidity assistance. The role of the central bank is in principle to remedy situations of illiquidity through emergency lending and even in this case, it should protect itself through taking appropriate collateral. If the bank that is in trouble is not viable, i.e., if its expected net future income falls short its net liabilities, the bank must be closed down. If the government considers that this would be too costly economically (in other words, that the bank's failure would exert strong negative externalities on the financial system and, through the credit it would no longer extend, on households and companies), it needs to provide it with fresh capital from budgetary resources. However, in practice, this distinction is difficult to make in real time, as illustrated by the Lehman Brothers case (see chapter 8). As a consequence, the central bank can find itself de facto engaged in the bailing-out of unviable institutions. The problem here is microeconomic in essence but it can acquire a macroeconomic dimension if many financial institutions provide low-quality collateral in exchange for central bank money.
- *Compatibility with price stability.* Central banks like to consider that the provision of liquidity in times of stress does not need to conflict with their macroeconomic objectives and in particular their price-stability mandate. This is certainly true when assistance is provided to one particular institution, but less so when they engage in wholesale liquidity

13. This does not need to be the case. In principle, the quality of assets is taken into account through applying "haircuts."

provision like in the aftermath of the crisis of 2007–09, or in the case of a small country whose banks are engaged in cross-border lending. In such situations, loans to banks result in an increase of the quantity of money that could result in inflation if extended beyond the liquidity stress period. However, even in such cases money creation is not necessarily permanent. At the end of a repo operation, banks hand back the amounts borrowed to the central bank, which destroys them. In order to avoid inflationary consequences, the central bank must gradually phase out its support and decrease the amount of its tender operations accordingly in order to remove the exceptional quantity of money brought to the market.[14] This applies even more to more extensive central bank intervention as discussed in chapter 8.

- *Implications for the definition of central bank objectives.* Central banks monitor asset prices as these convey information on possible future crises as well as on possible developments in inflation. In particular, a rise in asset prices may lead to imprudent borrowing and their eventual fall may result in financial disturbance. However, prior to the 2007–09 crisis there was no consensus on whether this was a justification for the central bank to include asset prices among its objectives. On the one hand, the Japanese experience called for such an inclusion: The expansionary monetary policy of the late 1980s was viewed in retrospect as a major cause of the asset-price bubble and the ensuing long stagnation of the Japanese economy; the bubble's burst pushed the banking sector into a severe crisis that had a very negative impact on output during the whole decade (figure 4.8). On the other hand, it was widely accepted that the central bank had no particular expertise for deciding on whether, for example, stock or housing prices are "'too high." The traditional response by central banks was therefore to discard asset-price stability as an objective and, according to the so-called *Greenspan doctrine*, to get ready to act aggressively in the event a bubble bursts (Greenspan, 1999; Bernanke and Gertler, 2001)

None of these three issues can be considered to be settled definitively. The role of central banks was once defined in a context where commercial banks were the main actors in the collection of savings and the allocation of financial resources, but traditional models are being challenged by the development of market-based finance, disintermediation, and the development of financial innovation. As central banks learn from experience and adapt to the new reality, responses to policy dilemmas are being reexamined. We return to this discussion in section 4.3.

14. See, for example, the article on the Eurosystem's open market operations during the period of financial market volatility in the May 2008 issue of the *ECB Monetary Bulletin*.

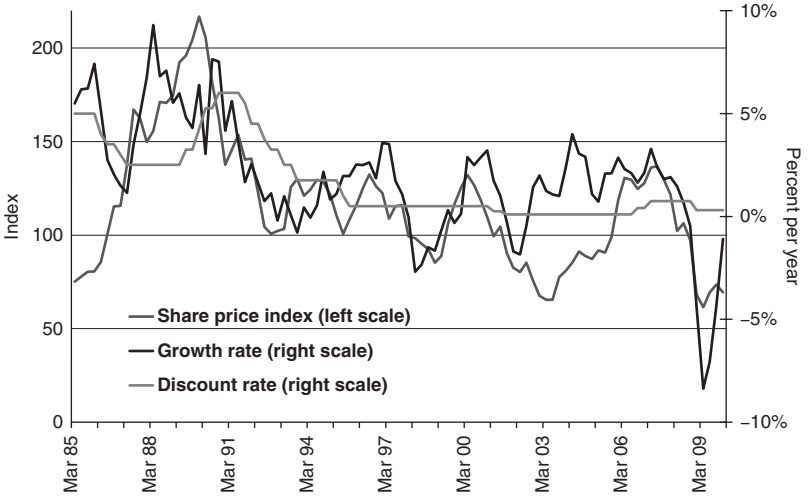


Figure 4.8 Interest rates, financial asset prices and growth in Japan, 1985–2009. Source: OECD.

Box 4.6 Interest Rates and the Pricing of Assets

Banks receive short-term deposits from their customers and hold long-term assets. These assets are loans to corporations and households as well as bonds, equities and real estate.

The market price of a bond is inversely related to the interest rate, for the following reason. Suppose a perpetual bond costs \$1 at time t and yields a 4% annual return, i.e., each year the holder of the bond will receive a 4 cent coupon. Suppose that, at time $t + 1$, interest rates have risen from 4 to 5%. This means that new bonds issued in $t + 1$ yield a 5% coupon. Nobody wants to buy the old bond unless it is cheaper. Its price thus falls until it reaches a value P such that the bond yields an intrinsic return of 5% despite paying coupon equal to 4% of the bond face value. This requires that $1 \times 4\% / P = 5\%$, i.e., $P = 0.80$: A one percentage-point rise in the interest rate triggers a 20% fall in the bond price. For bonds with finite maturities, the relationship is less straightforward but still exists. And it can be shown that the longer the maturity, the higher the sensitivity of the bond price to interest-rate changes.[a]

A similar, inverse relation between interest rates and the asset prices also holds for equities, but in a less mechanical way. The fundamental value of a stock is the price at which the investor is indifferent between, on the one hand, holding the stock and cashing in the dividends attached to it, and, on the other hand, selling it at market value. If investors are risk-neutral, the fundamental value is equal to the net present value of

expected future dividends. When the interest rate r is constant and the growth rate of dividends d_t is g, the price p_t of the stock is given by the *Gordon–Shapiro** formula:

$$p_t = \frac{d_t}{r - g} \tag{B4.5.1}$$

A higher interest rate r discounts more heavily future cash flows and therefore immediately lessens the value of the stock. In addition, the interest rate may affect the dividend through the macroeconomic equilibrium. In some cases, a decrease in r may increase g and magnify the stock price increase.

[a]It can be shown that the sensitivity of bond prices to interest-rate changes is equal to the *duration** of the bond, defined as the average date when investors will receive cash flows, each date being weighted by the size of the corresponding cash flow.

e) Summing up

Of the four objectives we have mentioned—price stability, exchange-rate stability, output stabilization, and financial stability, only the first one is formally included in all central banks' mandates. Financial stability is a core objective of most central banks, though not necessarily explicitly. The other objectives may or may not feature among the goals of the monetary institutions (table 4.1 and box 4.7).

Box 4.7 The Mandates of Four Central Banks

Through the Humphrey–Hawkins Act of 1978, the US Congress has assigned to the *Federal Reserve* the objective to "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." Furthermore, the Federal Reserve is entitled to provide emergency lending to banks.

As regards the *European Central Bank*, the EU treaty states that "The primary objective of the ESCB [European System of Central Banks] shall be to maintain price stability. Without prejudice to the objective of price stability, the ESCB shall support the general economic policies in the Community with a view to contributing to the achievement of the objectives of the Community as laid down in Article 3 [of the Treaty]" (article 127 TFEU). The objectives of the Community, such as stated in Article 3, are to "promote economic and social progress and a high level of employment and to achieve balanced and sustainable development." Financial stability is not explicitly part of the ECB mandate. The European

Table 4.1
The mandates of four central banks

	Legal vehicle	Price stability	Exchange-rate stability	Output stabilization	Financial stability
US Fed	Full Employment and Balanced Growth Act, 1978, a.k.a "Humphrey–Hawkins Act"	Yes	No, but may intervene on exchange markets, at the request of the US Treasury	Yes, on an equal footing with price stability	Yes
ECB	EU Treaty (since Maastricht Treaty of 1992)	Yes	No, but exchange rates are part of the second pillar of the monetary-policy strategy, and the ECB has the sole right to conduct foreign-exchange operations.	No, but may intervene on exchange markets	Not explicitly
Bank of England	Bank of England Act, 1998	Yes, definition of price stability belongs to government	No	Yes, secondary to price stability	Yes
Bank of Japan	Bank of Japan Law, 1997	Yes	No, but may be instructed to intervene on exchange markets	No, only as a consequence of price stability	Yes

System of Central Banks contributes to national policies with respect to financial stability (article 127 (5)), and the European Council may task the ECB with a supervisory role on banks (article 127 (6)) but this has not been decided so far.

The 1998 *Bank of England* Act gives it the mandate to "maintain price stability, and subject to that, to support the economic policy of the government, including its objectives for growth and employment." However, the UK Treasury may specify in writing "what price stability is to be taken to consist of" and it actually defines the Bank's price-stability objective. The Bank therefore is independent in fulfilling its mandate but is not free to decide how the mandate should be interpreted. Furthermore, it must report in writing in case it does not meet its inflation target. In addition to price stability, the Bank has a second core objective, financial stability, but responsibility in this field is shared with the Treasury and the Financial Services Authority (FSA). In 2007 the depositors' run on Northern Rock, a bank exposed in mortgage lending, exposed the ambiguities and the fault lines of this tripartite division of labor and prompted a rethink of the principles of financial regulation (see the Turner Review, 2009).

The 1997 *Bank of Japan* Law states that "currency and monetary control shall be aimed at achieving price stability, thereby contributing to the sound development of the national economy." The Bank of Japan is autonomous, but the law states that the Bank shall "always maintain close contact with the government and exchange views sufficiently." Financial stability and the ability to act as lender of last resort are explicitly part of the Bank's mandate.

4.2 Theories

Monetary-policy theory has been and still is a very active field of research, one of those where the dialogue between theoreticians and practitioners has been the most vibrant and one of those where theory has had major influence on the design of policy institutions.[15] In the 1960s and 1970s, the monetarist challenge to conventional Keynesian wisdom emerged from what was initially a critique of monetary-policy practices. Similarly, the rational-expectation models, which would have a profound impact on macroeconomic thinking and policy (see chapter 1), were initially developed in that context. The notions of time consistency and credibility, which would make their way

15. It is not by accident that central bank governors often have an active economic research background. Benjamin Bernanke (US), Stanley Fischer (Israel), José de Gregorio (Chile), Mervyn King (UK), Anastasios Orphanides (Cyprus), and Axel Weber (Germany) are all respected academics. In 2010, all ECB Board members but the President, Jean-Claude Trichet, had a PhD in economics.

into the basic toolkit of policymakers, were also first experimented within the monetary-policy field. Finally, the contemporary micro-founded neo-Keynesian models embodying price rigidities were developed in response and with the aim of providing sound theoretical foundations to monetary stabilization.

We start this section with a discussion of the principles that underpin monetary policy. We then move on to assessing its main transmission mechanisms, first in a closed- and second in an open-economy context. We end with a short discussion of the theoretical foundations of financial stability.

4.2.1 Principles

a) The long-run neutrality of money

The most fundamental question is whether monetary policy affects real variables. It is now widely accepted that changes in money supply do not affect real variables in the long run, a property known as the *long-term neutrality of money**. This dichotomy between money and real variables, which was first formalized by Scottish philosopher David Hume in 1742, is a consequence of the role of money as a unit of account: In the long run, doubling the quantity of money in circulation, or replacing a currency by another one of higher value, has no impact on real variables such as GDP, real wages, real interest rates, or the real exchange rate. Only nominal variables (nominal GDP, nominal wages, nominal interest rates, and the nominal exchange rate) are affected.

Hume's *quantity theory of money** is the simplest model consistent with this approach. It states that output is supply-determined and that the value of the transactions that can be carried out with one unit of money during a given period—the *velocity of money**—is exogenous. In this setting, there is a one-to-one relation between money growth and inflation. Controlling money growth allows the central bank to control the inflation rate without incurring any real cost (box 4.8).

Box 4.8 The Quantity Theory of Money

Money velocity V is defined as the nominal production allowed by the circulation of one money unit during one year:

$$PY = MV$$

where P denotes the general price level, M money supply, and Y real GDP. Assume Y grows as a constant rate as a consequence of population and productivity growth. Assume V is constant or evolves at a constant rate independently of monetary policy. If the central bank is able to control the growth rate of money supply, then, for a given GDP growth rate

$\Delta Y/Y$and a given evolution of velocity $\Delta V/V$, it is also able to control inflation:

$$\frac{\Delta P}{P} = \frac{\Delta V}{V} + \frac{\Delta M}{M} - \frac{\Delta Y}{Y}$$

According to this approach, the definition of a monetary-policy target requires estimating potential-output growth and the trend evolution of monetary velocity. The monetary target then follows.

In the tradition of the *Bundesbank*, the ECB in 1999 drew on the quantity theory of money to define the "first pillar" of its monetary strategy. This consisted in targeting money-supply growth at 4.5% a year, consistent with a 1.5% inflation rate, a 2.5% real GDP growth in the euro area, and a decrease of velocity by 0.5% a year:

$$1.5\% = -0.5\% + 4.5\% - 2.5\%$$

In such an approach, the monetary aggregate plays the role of an intermediate objective that is readily observable and more directly under the control of the central bank than the final objective of price stability, yet whose evolution is a good predictor of the final objective.

In 2003, the ECB decided to downplay this first pillar because money growth had been continuously higher than the target, without any major consequence for inflation (figure B4.8.1a). It has, however, not renounced monitoring of monetary aggregates (see section 4.3). It should also be noted that the link between money and inflation remains robust in high-inflation countries (figure B4.8.1b).

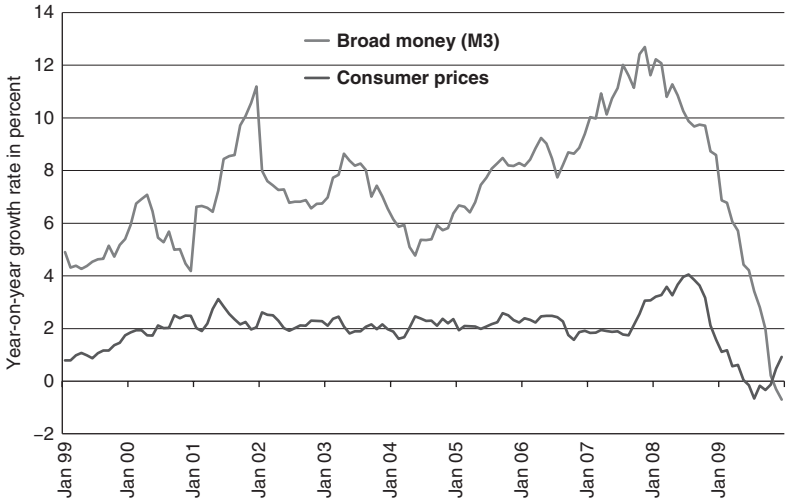


Figure B4.8.1a Money growth and consumer-price-index (CPI) inflation. Euro area, 1999–2009.

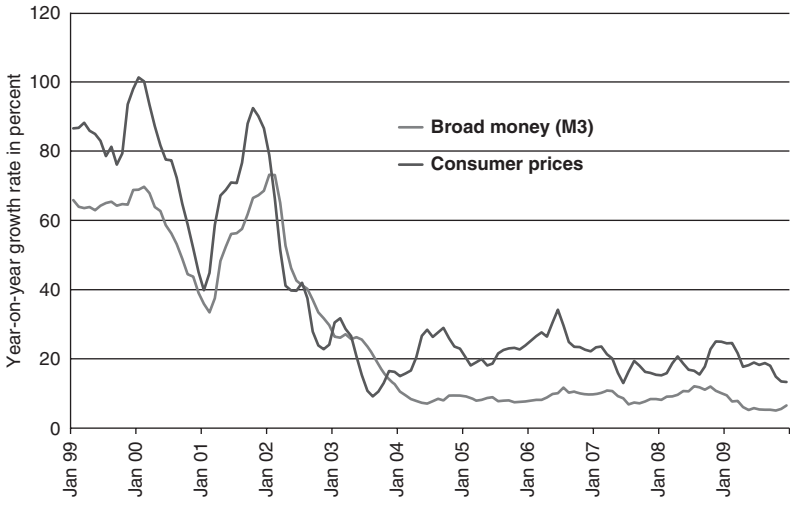


Figure B4.8.1b Turkey, 1999–2009.
Sources: a) European Central Bank, b) OECD.

As a cross-country, long-run regularity, the link between money growth and inflation—a consequence of money neutrality—raises little discussion. It has been documented in several studies (see, for example, McCandless and Weber, 1995, and Robert Lucas' Nobel lecture, 1996).

Two important caveats should be added, however. First, the neutrality of money does not imply that monetary policy has no influence whatsoever on real economic performance. In particular, high and unstable inflation is widely accepted as having detrimental effects on growth, as documented for example by Barro (1995), who finds that a 10 percentage point increase in the inflation rate results in a 0.2 percentage point reduction in the growth rate.[16] Second, the strength of the relationship between money growth and inflation comes from the long horizon and from the inclusion in the sample of high-inflation countries. In the short run and in a low-inflation context, there is little relationship between money growth and inflation, as illustrated in figure B4.8.1a of box 4.8. In the euro area, the high growth rate of M3 in the 2000s was accompanied by subdued headline inflation—hardly more than 2% per year.[17]

16. In other words one should distinguish between (a) the independence between the level of nominal variables, including the money stock, and real variables, and (b) the independence between the rate of change of nominal variables and that of real variables. The first proposition, known as the neutrality of money, is widely accepted, whereas the second, known as the *superneutrality of money**, is not.

17. Assenmacher-Wescher and Gerlach (2006) attempt to reconcile this observation with the quantitative theory of money by showing that different determinants affect inflation at different frequencies, and that the impact of money growth is only a long-run phenomenon.

b) Short-run nominal rigidities

One major explanation for the short-run disconnect between monetary growth and inflation is the existence of *nominal rigidities**, i.e., the fact that following a shock on the supply of money, prices and/or nominal wages adjust less than fully in the short run. Accordingly, a rise in money supply increases the real value of monetary holdings, which affects other real variables, including the real interest rate and real consumption.

In Keynes's *General Theory* (Keynes, 1936), a rise in money supply leads in the short run to a fall in the interest rate. This is because such a fall is the only way to raise money demand if prices do not adjust upward. A lower nominal and real interest rate encourage private agents to hold money balances in spite of their yielding no or little return and stimulates the demand for goods and services (which also in turn increases money demand). If there is excess production capacity, GDP rises. In the longer run, however, prices increase, which brings the interest rate and GDP back to their initial values, consistent with the long-run disconnect between nominal and real variables. Hence, in the Keynesian framework, money-market equilibrium is is achieved in the short run through nominal and real interest-rate adjustment rather than through price adjustment. Consistent with this determination, the saving-investment balance is achieved through output adjustment: If saving exceeds investment *ex ante*, total aggregate demand (consumption and investment) lies below aggregate supply and output will decline to meet the level of aggregate demand.

In brief, the short-run impact of monetary policy on real variables such as output or employment relies on incomplete price adjustment. Three types of explanation of nominal rigidities have been proposed in the literature: Imperfect information, staggered contracts, and menu costs.

- The imperfect-information theory of nominal rigidities was developed in parallel by Edmund Phelps (1967) and Robert Lucas (1972). It claims that producers have more information on their own prices than on those of the economy as a whole. Therefore, they can confuse a rise of the general price level with a price increase of their own product. In such a case, a monetary expansion initially leads to a rise in goods supply, until producers realize their mistake. This model played an important role in the 1980s by showing that nominal rigidities need not be explained by ad-hoc hypotheses on price adjustment and that they are compatible with rational behavior. However, it is hardly realistic as the general source of nominal rigidity. Indeed, it rests on a strong assumption concerning imperfect information and high price elasticity of supply. All in all, it can only apply to unexpected monetary policies.
- The second explanation of nominal rigidities, suggested by John Taylor (1980) and Stanley Fischer (1977), notes that contracts between firms and employees specify wages and possibly the conditions for their increases. The existence of such contracts is justified by the transaction

costs (including strikes) which would result from day-to-day adjustments of wages to market conditions. This model is realistic in that wage negotiations only occur by intervals, for instance once a year or at even lower frequency (traditionally every three years in US manufacturing industry). The macroeconomic implication is that wages do not react immediately to shocks. This creates nominal rigidity and in an economy in which, with staggered wage adjustments (at each period, only a fraction of wages adjust), monetary shocks have an impact on output even when they are perfectly anticipated.

- The third explanation of nominal rigidities is based on so-called "new Keynesian" models developed by George Akerlof and Janet Yellen (1985) and by N. Gregory Mankiw (1985), which try to reconcile the Keynesian theory with rational individual behavior. Their starting point is the finding that companies generally adjust their prices infrequently (once or twice a year), while the economic conditions that they confront (such as raw material prices, exchange rates, interest rates, etc.) evolve much more quickly. The response of the new Keynesian theory is that such behavior is optimal for each firm due to the existence of adjustment costs—called *menu costs** in reference to the printing costs incurred by restaurants when changing their menus. From a macroeconomic standpoint, however, such behavior is sub-optimal, because each firm fails to take into account that the delay in adjusting its price level temporarily contributes to creating an imbalance between supply and demand. In the case of a fall in demand, for instance, firms overlook the fact that a fall in the general level of the prices would support demand by raising the purchasing power of the money holdings. This lack of coordination leads the price level to adjust less than required.[18] Though the nature of menu costs is not perfectly clear, they seem to be relevant empirically.

The introduction of euro-denominated notes and coins on 1 January 2002 was a natural experiment in price setting. On a single day, all retail prices were redenominated in the new currency. For firms which had refrained from adjusting their prices in the previous months, this provided a perfect opportunity to make the adjustment without incurring an additional cost, because they had to redenominate all their prices anyway.

In their study of the effect of the euro on prices, Angeloni et al. (2006) find a one-off effect of the euro changeover on prices in the first quarter of 2002, consistent with the 0.2 percentage points increase documented by the European Statistical Office. This evidence is consistent with the menu-costs theory. It should be noted that consumers in the euro area overwhelmingly claim that the euro has had a strong one-off effect on prices, much beyond what statisticians and economists have found. One possible explanation would

18. See Calvo (1983).

be that consumers attach more psychological weight to frequent, small-sized transactions, for which rounding effects have been proportionally stronger, but which have a small weight in the aggregate price index.

The existence of short-term nominal rigidities is not incompatible with the long-term neutrality of money. A monetary expansion will have an impact on real variables in the short run, but this effect will gradually be phased out by price adjustment. Higher money growth may speed up price adjustment, because the cost of nonadjustment is greater. In the extreme case of hyper-inflation, price adjustment is almost instantaneous.

c) Optimal interest-rate setting

We have indicated in section 4.3.1 that the central banks' main monetary responsibility is to decide on the level of their interest rate(s). But what should guide this decision? In the 1960s the response to this question was largely ad hoc and discretionary. Then came the monetarist revolution of the 1970s and the 1980s, which advocated setting interest rates at a level consistent with the desired path for the monetary aggregates. However, as already mentioned, the link between money growth and inflation has proved to be loose, at least in the short run. In addition, financial liberalization and financial innovations have made the control of monetary aggregates difficult. Consistently, central banks have started looking for an alternative strategy. In response, new models of monetary policy have been developed in which monetary aggregates play a secondary role, or are altogether ignored.

This is the case with the model proposed by Richard Clarida, Jordi Galí and Mark Gertler (1999), which develops a "new Keynesian" theory of monetary policy (see box 4.9). In this model, the central bank sets the short-term interest rate so as to keep the future inflation rate and the future output gap as close as possible to its targets. Different weights can be given to the two objectives depending on the mandate of the central bank. The optimal level for the output gap is zero, which corresponds to a situation in which actual output equals potential output. In the model, optimal inflation is also assumed to be zero, but this is only for the sake of simplicity; the inflation target can be set at any constant level without changing the results.

An important aspect of the model is that the central bank is supposed to adopt a forward-looking approach. It does not attempt to control the current inflation or output gap but only to control their expected values. In a way, its true objectives are the forecasts for inflation and for the output gap. This is because delays in the monetary-transmission mechanisms do not allow the central bank to control current variables. This is an important distinction to keep in mind, and one that matters for discussions on monetary strategies.

Since the output gap (see chapter 1) is negatively related to the real interest rate and positively related to inflation, the two objectives of the central bank are consistent in the presence of demand shocks (which move inflation and the output gap in the same direction) but contradictory in the presence

of cost-push, or supply shocks (which move them in opposite directions). The policy implication of this observation is that the central bank should completely offset demand shocks even if it only cares about inflation, whereas it should only partially offset cost-push shocks. In early 2008, the ECB faced such a dilemma with rising inflation (due to oil and food price hikes) and declining output. Its initial response was to keep interest rates almost constant, which failed to stabilize both prices and the output gap. A few years before, in 2001, the ECB had not hesitated in cutting interest rates when output growth was declining in a context of low inflation, as this did not involve any policy dilemma.

Another implication of the model, which relies on rational expectations combined with auto-correlated shocks, is that the central bank should raise its interest rate by more than one percent when expected inflation increases by one percentage point, in order for the real interest rate to rise. This rule has been followed by the Fed and by the ECB since 1999.

Box 4.9 The "New Keynesian" Model of Monetary Policy (Clarida et al., 1999)

The model relies on two equations: An amended IS curve (B4.9.1), and an amended Phillips curve (B4.9.2):

$$x_t = -\varphi(i_t - E_t\pi_{t+1}) + E_t x_{t+1} + g_t \qquad \phi > 0 \qquad \text{(B4.9.1)}$$

$$\pi_t = \lambda x_t + \beta E_t \pi_{t+1} + u_t \qquad \lambda, \beta > 0 \qquad \text{(B4.9.2)}$$

where x_t denotes the output gap (i.e., the log-difference between actual and potential output) at time t, π_t is the inflation rate defined as the percentage variation of prices between $t-1$ and t, i_t is the short-run interest rate, E_t is the expectation operator, g_t is a demand shock and u_t a cost-push (supply) shock. Both shocks are assumed to be auto-correlated with autocorrelation coefficients equal to ϕ and ρ, respectively (this allows taking into account shocks that have some persistence and vanish only gradually over time).

Equation (B4.9.1) derives from optimization behavior of households with rational expectations. The presence of the expected output gap in this equation comes from consumption smoothing, and the real interest rate has an additional intertemporal substitution effect: When confronted with shocks, households tend to equalize consumption over time but they save more when the interest rate is high. Equation (B4.9.2) derives from staggered nominal price setting by monopolistically competitive firms: Because of short-run nominal rigidities, at each period a firm only has a given probability of being able to adjust its price at the level corresponding to profit-maximization. The price the firm chooses then depends on its expectation of future prices and on a discount factor β. The more nominal the rigidity, the less inflation depends on the current output gap (the lower λ).

Equation (B4.9.2) can be referred to as a Phillips curve because, for a given price expectation, it results in an upward-sloping relationship between inflation and output (or, equivalently, in a downward-sloping relationship between inflation and unemployment). This is like the original Phillips curve presented in chapter 1—however, in the short run only. In the long run, expectations adjust and the trade-off vanishes.[a]

Forward iteration in (B4.9.2) yields:

$$\pi_t = E_t \sum_{\tau=0}^{\infty} \beta^{\tau} \left[\lambda x_{t+\tau} + u_{t+\tau}\right] \tag{B4.9.3}$$

Inflation at time t depends on the whole sequence of expected output gaps and cost-push shocks from t to infinity. The central bank sets the nominal interest rate so as to minimize a loss function of the type presented in chapter 1:

$$Min\ L_t = \frac{1}{2} E_t \left(\sum_{\tau=0}^{\infty} \beta^{\tau} \left[\alpha x_{t+\tau}^2 + \pi_{t+\tau}^2\right] \right) \quad \alpha > 0 \tag{B4.9.4}$$

subject to (B4.9.1) and (B4.9.2), where α is the weight given to the output stabilization objective in comparison to the inflation objective. Equation (B4.9.4) implies that the central bank sets its interest rate at time t in order to keep the future output gap and inflation rate as close as possible to their target level, taking expectations as exogenous.

Since the output gap and inflation at date t do not depend on past values, but only on expected future values, the optimization problem can be solved as a succession of static decisions:

$$Min\ L_t = \frac{1}{2}[\alpha x_t^2 + \pi_t^2] + F_t \tag{B4.9.5}$$

subject to:

$$\pi_t = \lambda x_t + f_t \tag{B4.9.6}$$

with

$$F_t = \frac{1}{2} E_t \left(\sum_{\tau=1}^{\infty} \beta^{\tau} \left[\alpha x_{t+\tau}^2 + \pi_{t+\tau}^2\right] \right)$$

and

$$f_t = \beta E_t \pi_{t+1} + u_t$$

Since F_t and f_t can be considered exogenous by the central bank, the following first-order conditions hold:

$$\begin{cases} x_t = -\lambda q u_t \\ \pi_t = \alpha q u_t \\ i_t = \rho \alpha q u_t + \frac{1}{\varphi} g_t \end{cases} \quad \text{with} \quad q = \frac{1}{\lambda^2 + \alpha(1 - \beta\rho)} > 0 \tag{B4.9.7}$$

It is thus optimal for the central bank to raise the interest rate in the case of a positive demand shock or a positive cost-push shock. In the case of a demand shock, both the output gap and the inflation rate will remain *ex post* at their target levels because there is no contradiction between the two objectives. In the case of a cost-push shock, there is a contradiction between supporting aggregate demand and moderating inflation. Both targets cannot be reached simultaneously except if $\alpha = 0$ (the central bank does not target the output gap at all) or $\alpha = \infty$ (the central bank does not target inflation at all).

Since cost-push shocks are auto-regressive, a cost-push shock at time t will lead rational households to expect a positive inflation rate to persist at time $t + 1$ (if the autocorrelation coefficient ρ is positive):

$$E_t \pi_{t+1} = \rho \alpha q u_t \tag{B4.9.8}$$

Consequently, the reaction function of the central bank can be rewritten:

$$i_t = \gamma_\pi E_t \pi_{t+1} + \frac{1}{\varphi} g_t \tag{B4.9.9}$$

with

$$\gamma_\pi = 1 \frac{(1-\rho)\lambda}{\rho \varphi \alpha} > 1$$

This shows that the central bank must react to a one percentage point rise in expected inflation by increasing the interest rate by more than one percentage point.

[a]Technically the trade-off only vanishes entirely if $\beta = 1$, i.e., if the discount rate equals zero.

d) Central bank credibility

In the model presented in box 4.9, auto-correlated cost-push shocks, together with rational expectations, lead to expectations of inflation persistence and to a reaction by the central bank that raises the real interest rate. However, the optimal response to an adverse inflationary shock is to set the interest rate at the level that minimizes the loss to the central bank.

This result does not hold if the reason for inflation is that the central bank tries to push output above its natural level, i.e., push the output gap above zero. This is called an *inflation bias**. The problem was formalized by Robert Barro and David Gordon in an extraordinarily influential 1983 paper (chapter 2 and box 4.10). The parable told by Barro and Gordon

is a very simple one.[19] It starts from the assumption that the equilibrium output level is deemed too low by policymakers because it involves high unemployment, but that unemployment has in fact a structural character. If the central bank mistakenly targets a higher level of output in order to reduce unemployment, the outcome is bound to be inflationary because only structural polices (such as labor market reforms or tax reforms) can lower structural unemployment. As households are assumed to know the true economic parameters and the central bank's preferences, they will rationally expect inflation and efforts to reduce unemployment will be frustrated. Only inflation will remain.

In fact, the mechanism of the inflationary bias originates in the *augmented Phillips curve** theory introduced by Edmund Phelps (1967) and Milton Friedman (1968). As explained in chapter 1, the Phillips curve yields a negative relationship between the unemployment rate and the rate of change in nominal wages. However, this only applies *for a given expectation of inflation*. Phelps and Friedman argued that, at any unemployment rate, nominal wages would grow faster in response to a rise in the level of the expected consumer price inflation, because workers would be eager to defend the purchasing power of their income. This would in turn result in higher inflation and thereby in higher inflationary expectations. In the long run, it is reasonable to assume that nominal wages and prices grow at the same rate (but for the effect of technical progress) and the implication is that, whatever the inflation rate obtained in the long run, the unemployment rate will return to an equilibrium value called the *Non-Accelerating Inflation Rate of Unemployment* (*NAIRU*)*. In other words, the Phillips curve is vertical (figure 4.9). Hence, there is no trade-off between inflation and unemployment in the long run, and policies aiming at reducing the unemployment rate below the NAIRU only lead to higher inflation in the long run. The Barro–Gordon model formalizes the inflation bias when the central bank has an optimizing behavior, and Clarida et al. (1999) reconcile these ideas in a model with micro-foundation.

The inflation bias disappears if the central bank can commit to a certain inflation target—for instance, because it is independent with an explicit inflation-targeting mandate or because it is more inflation-averse ('*conservative**') than society. In this case, private agents will no longer anticipate an excess of monetary expansion, or a mitigated reaction to cost-push shocks. By reducing inflation expectations, such a strategy is designed to reduce the need for high interest rates in the short term. This in turn reduces the output cost of fighting inflation. For this to happen, the central bank needs to be regarded by the public as bound by its mandate or truly conservative.

Barro and Gordon's influence results from their having provided a simple but forceful case for central bank independence. Their paper was, however,

19. For a critical appraisal, see Blinder (1997). An academic, Alan Blinder also held the position of Vice-Chairman of the Fed.

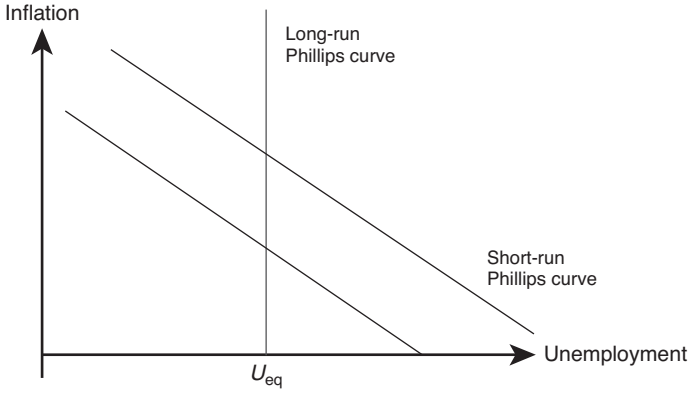


Figure 4.9 The Phillips curve.

part of a broader literature that emphasized central bank credibility. The *credibility** of the central bank can be defined as its ability to stick to its own policy announcements. Credibility is undermined by the *time-inconsistency problem* (chapter 2), i.e., the fact that a given announcement may be optimal today but no longer so tomorrow, which calls for re-optimization from the central bank. This is a perfectly appropriate behavior if departure from initial plans is justified by new, unexpected economic shocks. However, the central bank can also be tempted to cheat on citizens, i.e., to announce a low-inflation policy but to renege on it—for example to reduce unemployment. To the extent this behavior is understood by agents, they will no longer base their expectations on the central bank's announcement. For this reason, the central bank may need to tie its hands to a *monetary rule**. A popular rule is *inflation targeting**, where the central bank targets the average expected inflation rate over the next one or two years. Another one, especially in developing countries, is a fixed exchange-rate regime, where the central bank commits to intervene so as to keep the nominal exchange rate stable (more on exchange-rate regimes in chapter 5).

For the credibility of the central bank, it is also important that the central bank is able to embrace a long horizon. This is the case, for instance, if the same governors are in office when the policy is announced and when its outcome is observed. This justifies long mandates and independence from politicians bound by the election cycle.

To enhance credibility, most modern central banks combine a mandate to achieve price stability, formal independence from the government, long mandates, and a commitment scheme such as inflation targeting.

Another scheme, suggested by New Zealand economist Carl Wash (1995), is to design an incentive contract for the central bank governor and board members that makes the governor's compensation a function of the bank's inflation record. Walsh considers delegation of monetary-policy responsibility

to be a principal-agent issue, where the principal (the government) delegates to the agent (the central bank) the task of attaining a certain objective (low inflation). He shows that the inflation bias will be eliminated if the central banker receives compensation that negatively depends on money growth (considered as the monetary instrument) or inflation.

However, central bank credibility is a broader concept. Alan Greenspan, who chaired the Federal Reserve from 1987 to 2005, never endorsed a predefined rule, but while in office his personality and the track record he built were sufficient to ensure he had very high credibility. This credibility did not rely on institutional devices, but on *reputation**. What convinced agents that Greenspan would not try to fool them was that he would have paid the high price of losing all his accumulated reputation. Here again, the longevity of central bankers is crucial to their credibility.

Box 4.10 The Inflation Bias and the Conservative Central Banker (Barro and Gordon, 1983; Rogoff, 1985; Clarida et al., 1999)

The Barro–Gordon model presented in chapter 2 can be plugged into the framework of box 4.9. Assume that the central bank targets a positive output gap k. The loss function (B4.9.4) is modified accordingly:

$$Min\ L_t = \frac{1}{2}E_t\left(\sum_{\tau=0}^{\infty}\beta^{\tau}\left[\alpha(x_{t+\tau}-k)^2+\pi_{t+\tau}^2\right]\right) \qquad \text{(B4.10.1)}$$

For simplicity, β is assumed to be equal to unity, i.e., there is no discount of the future. The optimization program under the constraints (B4.9.1) and (B4.9.2) yields:

$$\begin{cases} x_t^{k>0} = x_t^{k=0} \\ \pi_t^{k>0} = \pi_t^{k=0} + \dfrac{\alpha}{\lambda}k \end{cases} \qquad \text{(B4.10.2)}$$

where the subscript $k = 0$ refers to the baseline solution (see box 4.9) and $k > 0$ to the solution when the central bank targets a positive output gap. From equation (B4.10.2), it follows that the central bank fails to increase output but does increase inflation. Clearly there is a loss in comparison to the baseline case. The positive inflation bias $\alpha k/\lambda$ positively depends on k, the desired increase in output, and on α, the preference for output stabilization. Appointing a conservative central banker who assigns a lower weight than society to output stabilization reduces the inflation bias (Rogoff, 1985).

Whatever the solution adopted to counter the inflation-bias problem (be it a monetary rule, the appointment of conservative central bankers, reputation,

or an incentive contract), achieving the goal is facilitated if the central bank is independent from political power. Independence consists in appointing central bankers for very long, fixed-duration mandates (except for serious misconduct); to prohibit any pressure from the governments; and to give to the central bank budgetary independence. There is some empirical evidence (Alesina and Summers, 1993) that over the long term, inflation is negatively correlated to the degree of independence of the central bank in industrialized countries.

For central banks with an explicit inflation target, the comparison between the target and inflation expectations over the medium term provides a good measure of credibility. As already noted in section 4.1, inflation expectations cannot be observed directly but can be inferred from surveys or by comparing the returns of inflation-indexed and nonindexed bonds.

e) Are monetary and fiscal policy interdependent?

In the long run, complete independence of monetary policy from fiscal policy is only possible if fiscal policy is sustainable or if the central bank is indifferent to the risk of government bankruptcy (box 4.11). If the public debt ratio exceeds its sustainable long-run level and fiscal authorities refrain from undertaking a fiscal retrenchment, asset holders will anticipate either government *default* (where creditors are not reimbursed) or *debt monetization* (where the central bank bails out the government through a massive purchase of its bonds and raises money supply accordingly). In the former case, the central bank may be hurt by the loss of value of its assets. More importantly it is likely to be wary of the economic consequences of commercial bank defaults. Hence it is likely that the central bank will prefer the latter case, monetization, with its inflationary consequences.

Box 4.11 Monetary Consequences of Deficits: The "Unpleasant Arithmetic" of Sargent and Wallace (1981)

Monetary and fiscal policies are normally regarded as mutually independent. Sargent and Wallace (1981) have shown that this is not true in the long run.

The starting point is an overlapping-generations model: At each period t, N_t young people and N_{t-1} old ones coexist, with $N_t = (1 + n)N_{t-1}$ where $n > 0$ is the growth rate of the population. When he or she is young, each individual receives an endowment and can decide to save part for old-age consumption. Savings are held in the form of money or public bonds. Public bonds bought at t entitle each individual to receive $(1 + R_{t+1})B_t/N_{t+1}$, where R_t denotes the interest rate. The government levies a tax τ on each young person, consumes $P_t G_t$ and finances the remaining deficit by debt and/or by monetary creation M_t. P_t is the price level.

The government's budgetary constraint is:

$$G_t + (1 + R_t)B_{t-1} = \tau N_t + B_t + \frac{M_t - M_{t-1}}{P_t} \qquad \text{(B4.11.1)}$$

Consider the steady state where the real amounts per capita ($b = B_t/N_t$, $g = G_t/N_t$ and $m = M_t/N_t$) are constant and where money supply grows at a constant rate $\lambda = M_t/M_{t-1} - 1$. By dividing the above formula by N_t, the government's fiscal constraint becomes:

$$g = \tau + \left(1 - \frac{1 + R}{1 + n}\right) b + \frac{m}{p}\left(1 - \frac{1}{1 + \lambda}\right) \qquad \text{(B4.11.2)}$$

If the real interest R rate is higher than the growth rate n, then debt dynamics are divergent (see chapter 3). A rise in public consumption Δg financed by an increase in debt Δb triggers an explosion of the debt, unless there is a rise in the tax burden τ or an increase in the growth rate of the money supply λ. By contrast, if $R < n$, the new debt is easily refunded and even makes it possible to lower the tax burden or to tighten monetary policy.

The long-run interdependence between fiscal and monetary policy implies that lasting monetary stability is very unlikely if the fiscal authority behaves in an irresponsible way. An example of this type was provided by Argentina prior to the 2002 crisis: Although the country was committed by law to maintaining a fixed exchange rate to the US dollar and the currency issued by the central bank was supposed to be fully backed by the foreign exchange reserves (this regime is called a *currency board*, see chapter 5), the profligate behavior of the federal and especially sub-federal fiscal authorities was never reined in. Ultimately, the government was forced to abandon the dollar peg and this led to a violent currency and financial crisis. Though the mechanism was not identical to that in the closed-economy setting of box 4.11, the logic was the same. In the euro area, the Greek crisis that broke out in 2010 was of the same nature.

Policy regimes where monetary policy is subordinated to the goal of assisting in the financing of government budget are generally called *fiscal dominance** regimes (Woodford, 2001). This is generally the case in wartime. In the US, a 1942 agreement committed the Federal Reserve to maintaining "relatively stables prices and yields for government securities." Up until the termination of this agreement in 1951, monetary policy was given the objective of keeping long-term interest rates low and the goal of maintaining price stability was assigned to price controls. Indeed from 1942 to 1947 at least, short- and long-term interest rates barely changed. This helped the US support the war effort without incurring the corresponding debt costs.

While this type of situation, though not uncommon in wartime, is very infrequent in peacetime, the consequences of monetary–fiscal interactions

also apply in a setting where the central bank aims at controlling inflation but not the price *level*. Woodford (2001) proposes a formalization where monetary policy follows a Taylor rule and is therefore able to avoid an inflation drift but where the price level is determined by the fiscal sustainability condition. This approach is known as the *fiscal theory of the price level**.

This long-run interdependence is the main justification for limiting public borrowing in a monetary union, as discussed in chapter 3.

In the short run, there is no consensus on the desirability of coordinating monetary and fiscal policies to achieve a *policy-mix**, at least as long as monetary policy remains effective. Advocates of this co-ordination generally put forward two arguments.

- The first is of a political nature: It is argued that governments and the central bank should jointly be responsible for macroeconomic management; otherwise each of them could be held responsible by an uninformed public for the errors made by the other.
- The second argument is an economic one: In the short run, monetary and fiscal policy both affect aggregate demand. In the absence of co-ordination, a noncooperative equilibrium could emerge between the government and the central bank, whereby each player attempts to take into account the other player's reaction to its own policy. In this equilibrium (a Nash equilibrium in game-theoretical terms) monetary policy will be too tight in order to compensate the excessively loose character of budgetary policy, and reciprocally budgetary policy will be too loose in order to compensate the high level of interest rates (Beetsma and Uhlig, 1999; Dixit and Lambertini, 2003). Coordination is expected to make it possible to reach the first-best equilibrium. However this equilibrium risks being unstable, as a deviation by one of the players leads the other one to revert to noncooperative behavior.

Opponents of coordination point out that coordination by nature threatens central bank independence and argue that the game-theoretical problem involved in the rivalry between monetary and fiscal policy can be solved by making monetary policy fully independent. In this kind of setting monetary policy can raise the interest rate without limit while fiscal policy action is limited by the public finance cost of deficits. Monetary policy therefore always "wins," which solves the problem.

Specific coordination issues arise when monetary policy reaches the zero bound on nominal interest rates and embarks on unconventional policies, as will be discussed in chapter 8.

4.2.2 Transmission channels

So far, we have only discussed *why* monetary policy can affect real variables. Here, we discuss *how* it impacts aggregate demand, starting with the closed economy. Three main *transmission channels** are generally distinguished:

The *interest-rate channel*, the *asset-price channel*, and the *credit channel*. All three obviously operate in parallel and contribute to the general equilibrium outcome, but distinguishing them helps understand how monetary policy works, and what determines the magnitude of its impact. It can also be interesting from a policy standpoint, especially because financial reforms affect the transmission of monetary impulses through each of the channels.

a) The interest-rate channel

The *interest-rate channel** is the traditional Keynesian channel: In the presence of nominal rigidities, a monetary expansion leads to a fall in the (nominal and real) interest rate, hence to a revival of investment and durable-goods consumption. In the short run, the rise in those categories of spending in turn results in a multiplier effect (see chapter 3) on the demand for goods and services.

Note, however, that the only interest rate which is directly affected by monetary policy is the overnight, nominal interest rate, while aggregate demand depends on expected real interest rates at longer-term horizons. The impact of a monetary-policy move thus depends on (i) which interest rates matter most for economic agents, and (ii) how these interest rates are affected by the change in the overnight rate. Evidence shows that countries differ considerably along the first dimension: For example, mortgage rates in the UK tend to be variable and indexed on short-term rates, which implies that monetary-policy decisions immediately affect both the cost of new borrowing and the disposable income of indebted households; in contrast, German households borrow fixed-term, which insulates them from monetary impulses once in debt. There are also differences along the second dimension: As explained in section 4.1, whether short-term rates affect long-term rates depends on expectations about the future monetary policy. The strength of the interest-rate channel therefore varies across countries.[20]

b) The asset-price channel

The *asset-price channel** relies on the negative relationship between asset prices and interest rates (see box 4.5 in section 4.1): A decrease in the interest rate generally raises the value of financial assets held by households, who, in turn, partially consume this extra wealth. Such wealth effects played an important role in Japan in the early 1990s, when the burst of the asset-price bubble had a negative impact on consumption; in 2001, the sharp fall in US stock prices also had a negative impact on consumption, whereas the rise in real estate

20. This has been a topic for research and policy discussions in the euro area as differences in borrowing practices imply asymmetries in the transmission of the same monetary impulse to member countries. See Angeloni and Ehrmann (2003).

prices tended to sustain US consumption during the 2000s. The asset-price channel also affects the corporate sector: A rise in stock prices increases the profitability of new capital expenditures (also known as *Tobin's q**),[21] which supports investment.

The importance of the asset-price channel has increased over time as a consequence of the general rise in the wealth-to-income ratio and the increased sophistication of financial markets which allow households to withdraw equity from their wealth without actually selling assets. In Anglo-Saxon economies, the so-called *mortgage equity withdrawal**(the difference between new housing finance and actual investment in housing) played an important role in supporting household consumption in the early 2000s. In the UK, for example, such annual withdrawal amounted on average to 6% of post-tax household income in the 2002–06 period and exhibited significant volatility.[22]

c) The credit channel

Finally, the *credit channel** results from the impact of the interest rate on the supply of—rather than the demand for—credit: In response to an improvement in their refinancing conditions, banks tend to increase their supply of credit.

The reason for this is a subtle one (Bernanke and Gertler, 1995). In an imperfect-information world, it is costly for banks to assess properly the quality of all the investment projects for which borrowers—especially for small- and medium-sized enterprises—request loans. Lack of information on the quality of projects forces them to include a default premium in the credit cost proposed to all companies—which penalizes or even dissuades good investment projects whose probability of failure is low. However, risky projects may not be discouraged, as borrowers know that their probability of failure is high and accept paying the corresponding premium. The more banks increase the interest rate, the more they actually discourage good projects and select bad ones (box 4.12). This adverse selection problem, very well known in insurance theory, leads banks to restrict credit rather than price risk.[23]

21. Tobin's q is the ratio of the market value of companies to the cost of renewal of their stock of physical capital. It is the central variable of the neoclassic theory of investment. When q increases, the market value of the company increases in relation to the replacement cost of the capital; therefore the price of new equipment falls relative to the cost of its financing through issuing shares, which leads to a rise in investment. In the absence of adjustment costs of the capital stock, one should permanently have $q = 1$; when the adjustment is not immediate, investment depends positively on q. Depending on models, investment depends on average q or on marginal q (the ratio of the incremental increase of the company's value and the cost of additional capital). See for example Caballero (1999).

22. See the Bank of England quarterly estimates on the Bank's Web site.

23. It is for the same reason that poor, single-person entrepreneurs are cut off from bank credit. The development of micro-credit can be regarded as a way of overcoming credit restrictions through a system of mutual screening and guaranteeing.

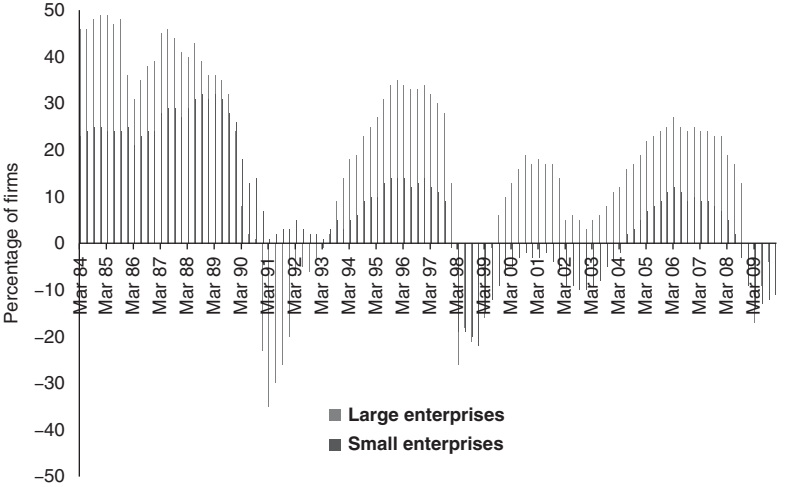


Figure 4.10 Credit discrimination in Japan, 1984–2009.
Source: Bank of Japan *Tankan* Quarterly Survey.
Reading: Difference between the proportions of firms facing "accommodative" and "severe" lending attitudes. Change in methodology in 2004.

Credit rationing especially affects small- and medium-size enterprises, since they do not have access to capital markets and depend on bank financing. An illustration of this phenomenon is the Bank of Japan "Tankan" survey which measures the lending attitude of Japanese banks (figure 4.10). The difference between the proportion of enterprises facing "accommodative" and "severe" lending attitudes fluctuates according to the business cycle, and small-sized enterprises always feel more constrained in their access to credit.

When the short-term interest rate decreases, the rational response of a profit-maximizing bank is to relax credit constraints—hence, an impact on credit supply that does not take the form of price changes. In addition, a lower interest rate also raises the value of the assets used to guarantee the loans, and therefore the companies' access to credit (Kiyotaki and Moore, 1997).

The banks' financial health is crucial for the transmission of monetary policy: When the banks' balance-sheets are burdened with *nonperforming loans**, i.e., loans with high probability of default, or with *impaired assets**, i.e., financial assets that are not traded any more or whose market value is much lower than when they were purchased by the bank, banks are less willing to grant new loans. This second source of credit rationing—often called *credit crunch**—was the main explanation for the poor effectiveness of Japanese monetary policy at the end of the 1990s and at the beginning of the 2000s. The Bank of Japan brought its leading rates nearly to zero in 1995 (see figure 4.4 in section 4.1) but with little effect. Even the adoption

in March 2001 of expansionary targets for the monetary aggregates remained without significant impact on credit and economic activity until the banks' finances were restored through recapitalization. The same happened in Sweden and Finland, which underwent a banking crisis at the beginning of the 1990s. In the US and in Europe in 2008, the deterioration in the quality of the banks' balance sheets also led to credit supply constraints which were initially obscured by companies drawing massively on credit lines banks had previously committed to extend to them (Ivashina and Scharfstein, 2010).

The link between monetary policy and fiscal policy therefore does not only run from the latter to the former, through debt monetization. Public money can also be crucially needed to restore the effectiveness of monetary policy, through a recapitalization of banks, and by relieving them of their impaired assets. This latter point has been forcefully put forward by the International Monetary Fund throughout the 2007–09 crisis.

Box 4.12 Credit Rationing (Stiglitz and Weiss, 1981)

In a seminal paper of 1981, Joseph Stiglitz and Andrew Weiss consider a bank loan as a contract between the bank and the borrower, defined by its amount, maturity, interest rate and collateral, i.e., an asset that the bank can seize if the borrower is unable to repay the loan.

As a general rule, the supply of credit by the bank is an increasing function of the interest rate, whereas the demand for credit is a decreasing function of the interest rate. Hence both sides will agree on an amount and an interest rate that correspond to the intercept of the two curves. However, Stiglitz and Weiss note, due to information asymmetry, the supply of credit may not be a monotonic function of the interest rate. Suppose there are two categories of borrowers: The safe (who are likely to repay their debts), and the risky (who are likely to default). The bank cannot observe whether a specific borrower is safe or risky (whereas the borrower knows his or her own type). Hence the bank applies the same interest rate to both types of borrowers. This rate includes a risk premium.

The higher the interest rate, the less inclined are safe borrowers to take a loan, because they know they will effectively have to pay the corresponding cost. Hence, a rise in the interest rate does not necessary increase the expected profit of the bank. The supply of credit $S(r)$ may correspondingly be an increasing function of the interest rate up to a certain threshold r^* but a decreasing function for $r > r^*$ (figure B4.12.1). A consequence is that the supply curve $S(r)$ may not cross the demand curve $D(r)$. In the figure, this is the case with $D_2(r)$. Then, borrowers are rationed, since they will not obtain the amount of credit they need—*for any interest rate.* This situation is called *credit rationing**.

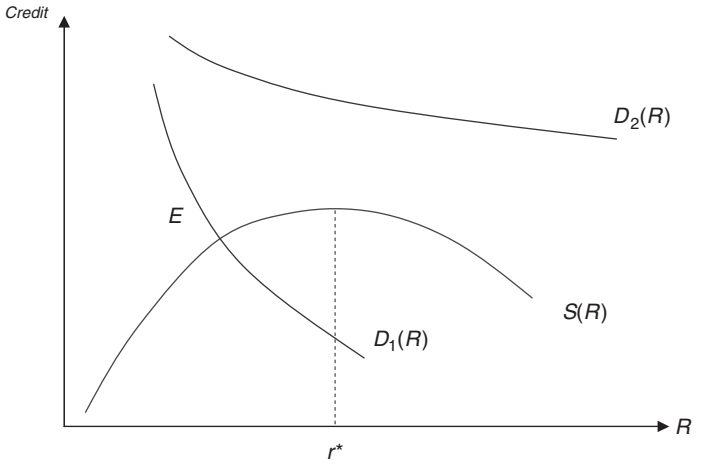


Figure B4.12.1 Credit rationing.

To avoid such situations, banks tend to structure their credit supply so as to force the borrowers to reveal their risk. For instance, the interest rate proposed will be lower the higher the collateral. This helps banks to assess the level of risk of each borrower and to propose an interest rate with the corresponding risk premium.

To avoid credit rationing, public authorities can also impose standards for the disclosure of information on nonfinancial companies, so as to allow lenders and equity investors to better discriminate between debtors. This is the purpose of financial reporting standards. Financial reporting procedures were tightened in the US after a series of bankruptcies in the early 2000s, most notably through the Sarbanes–Oxley act, and in the rest of the world through the adoption of the *International Financial Reporting Standards*.[24]

Public intervention can also consist in providing public support when the private yield of a project is insufficient, but the social yield is high enough to justify investment (e.g., for infrastructure projects). However, it must not be forgotten that credit rationing is a rational behavior of banks, which are unable to discriminate "good" risks from "bad" ones, and that systematically subsidizing "bad" risks is unproductive.

It can be noted that none of these transmission channels relies on a direct effect of money growth on inflation, as postulated by the quantity theory of money. In our set-up, the impact of money growth on inflation is channeled by interest rates, asset prices and bank credit through their respective influence on aggregate demand. A direct link between monetary policy and inflation could

24. On the economic consequences of accounting standards, see chapter 8.

Table 4.2
Predominant type of household mortgage interest rate

Australia	Variable	Italy	Mixed
Austria	Fixed	Japan	Mixed
Belgium	Fixed	The Netherlands	Fixed
Canada	Fixed	Norway	Variable
Denmark	Fixed	Portugal	Variable
Finland	Variable	Spain	Variable
France	Fixed	Sweden	Variable
Germany	Fixed	Switzerland	Variable
Greece	Variable	UK	Variable
Ireland	Variable	US	Fixed

Source: Debelle (2004).

be introduced by assuming that price expectations are affected by monetary policy. It would, however, be illogical to introduce expectations that are not consistent with the assumptions of the model.

d) Assessing the channels

The strength of the various transmission channels varies from country to country. The higher the proportion of short-term or variable-rate loans in the country, the stronger is the interest-rate channel. In Europe, this first criterion tends to indicate that the interest-rate channel is more powerful in the UK or in Spain than in France or Germany (table 4.2). The asset-price channel depends on the extent of asset holdings by domestic consumers. In the US, households can be especially reactive to it since they are both indebted at mainly variable rates and holders of large financial and real estate wealth. Finally, the importance of the credit channel depends on share of small-to-medium-sized enterprises (SMEs)in output and on their dependence vis-à-vis bank credit.

Econometric studies have tried to measure the impact of monetary policy in various countries. One popular methodology is the Vector Autoregressive analysis (VAR), which provides a comprehensive description of the impact of monetary policy, yet relies on few assumptions concerning the functioning of the economy and the transmission channels at work (box 4.13).

Box 4.13 Monetary-Policy Transmission Channels in Practice

One popular way of studying monetary-policy transmission channels is to estimate a dynamic econometric model describing the joint variations of GDP, prices, employment . . . and of one or more monetary instruments. Such a VAR model is written as:

$$X_t = A_1 X_{t-1} + \ldots + A_k X_{t-k} + u_t \qquad \text{(B4.13.1)}$$

Where X_t is the vector of the macroeconomic variables under review (in difference with a baseline path). For instance, $X_t = (i_t, y_t, p_t)'$ with i_t the nominal interest rate, y_t the logarithm of real output, and p_t the logarithm of the consumer-price index. A_j ($j = 1$ to k) is a 3×3 matrix of the coefficients to be estimated, and u_t is the vector of the three residuals at time t. The fact that i_t is exogenous will be reflected by the fact that the first line of all matrices A_j is (1 0 0). This hypothesis can be imposed or tested statistically.

It is then possible to calculate the *impulse-response function** that shows the dynamics of X_t over time following a given innovation u at time zero. In the present example, the impulse response function of real output $(y_1, y_2 \ldots)$ to a temporary unitary shock on the first-period nominal interest rate i_0 is $(\zeta_1, \zeta_2 \ldots)$ where:

$$\zeta_j = (0\ 1\ 0) \sum_{l=1}^{j} A_l (1\ 0\ 0)' \qquad \text{(B4.13.2)}$$

In the same setting, the response of real output at time j to a permanent shock on the nominal interest rate is $\sum_{i=1}^{j} \zeta_i$.

Figure 4.11, which is based on such methodology, gives the impact of the same restrictive monetary policy in the euro area, the US, and the UK. As expected, output falls after one quarter and prices also fall but after a longer time. Strikingly, though, the reaction of the economy is larger and more persistent in the US and UK than in the euro area. GDP returns to its initial level after approximately eight quarters in the euro area, 12 quarters in the US and 16 quarters in the UK.

4.2.3 Monetary policy in an open economy

a) Monetary conditions

A fall in the domestic interest rate implies a fall in the yield of other domestic assets that can be substituted for bonds. To the extent that they are not impeded by capital controls, domestic and foreign asset-holders are discouraged from holding domestic assets and encouraged to hold foreign assets. Consistent with the law of supply and demand, the relative price of domestic assets must fall. In a floating-exchange-rates context, this amounts to a *depreciation* of the currency. Conversely, an increase in the domestic interest rate triggers a currency *appreciation**. In turn, the variation of the exchange-rate influences: (i) The price level through the implied change in import prices; (ii) aggregate demand through substitutions between traded and nontraded domestic goods and between domestic and foreign-traded

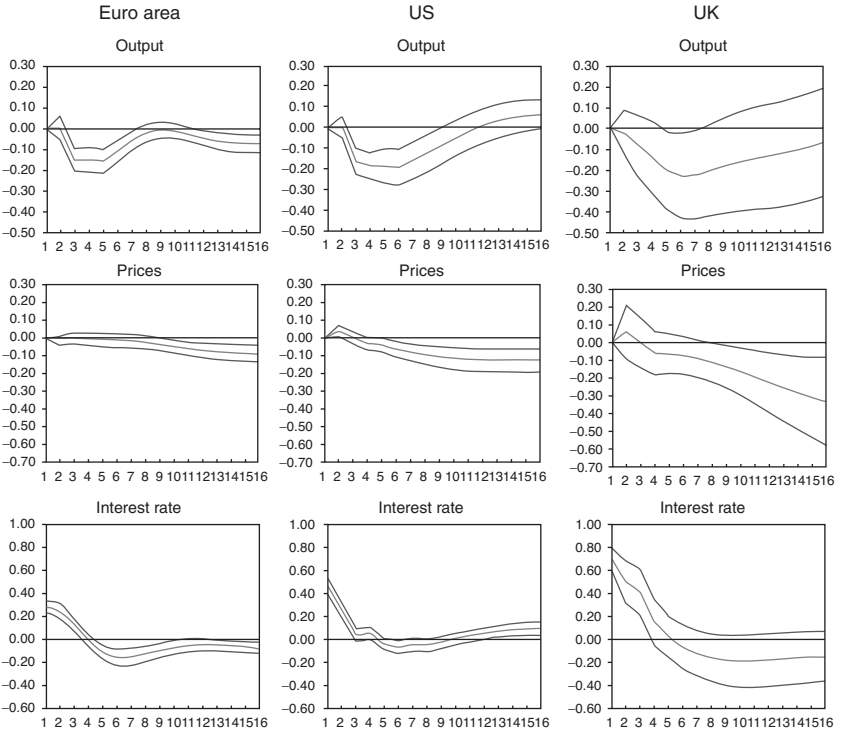


Figure 4.11 Impact of a restrictive monetary-policy shock in a VAR model.
Source: Bean et al. (2002).
The horizontal axis shows the number of quarters after the initial shock and the vertical axis differences with the baseline on percentage points.

goods; and (iii) aggregate supply through a change in the relative price of inputs. Through this mechanism, the *external channel** of monetary policy amplifies the effect of monetary policy.

The external channel is important in small, open economies where exchange-rate changes play a major role in the determination of prices and output. This implies that interest rates alone cannot be a sufficient indicator of the stance of monetary policy and calls for a broader approach. Central banks can use a *monetary conditions index** (MCI), as initially proposed by the Bank of Canada. A MCI is a weighted average of the (nominal or real) interest rate and the (nominal or real) exchange rate, with weights reflecting their respective roles in the determination of aggregate demand. The usefulness of the MCI hinges on the existence of a stable relationship between the exchange rate and output. Furthermore, it is not a measure of the impact of monetary policy, because, while the policy rate is under the control of the central bank, the exchange rate can float freely. The MCI can therefore vary in the absence of any monetary policy change.

b) The impact of the interest rate on the exchange rate

Assuming perfect capital mobility and no risk aversion, an investor will invest his or her wealth in the currency with the highest expected return. Denoting i the domestic interest rate, each unit of domestic currency invested domestically will be worth $(1+i)$ units after one year. Similarly, if i^* represents the foreign interest rate, each unit of foreign currency invested abroad today will be worth $(1+i^*)$ units of foreign currency after one year. However, to invest abroad, the domestic investor needs first to change his or her wealth into foreign currency. Denoting S the exchange rate (the number of foreign currency units per unit of domestic currency), one unit of domestic currency allows the investor to buy S units of foreign currency, which will be worth $S(1+i^*)$ after one year. He or she will then be able to convert this amount back at the expected exchange rate S^e. Hence the repatriated amount in domestic currency is expected to be $S(1+i^*)/S^e$.

A risk-neutral investor will invest wherever the expected return is higher. This implies at equilibrium:

$$(1+i) = \frac{S(1+i^*)}{S^e} \tag{4.2}$$

This equality is called the *uncovered interest parity* (*UIP*)* and it must hold in the absence of capital controls if domestic and foreign assets are perfectly substitutable and if investors are risk-neutral.[25] Considering relatively small interest rates, and noting $s = \ln(S)$ and $s^e = \ln(S^e)$, the linearized form of the uncovered interest parity, which is true for small variations of S and i, is:

$$i = i^* - (s^e - s) = i^* - \Delta s^e \tag{4.3}$$

where Δs^e is the expected appreciation of the domestic currency ($\Delta s^e = s^e - s$). For instance, suppose that the national interest rate i is 3% while the foreign interest rate i^* is of 4%. Assuming they would not anticipate any variation of the exchange rate ($\Delta s^e = 0$), residents would invest abroad; this would make s fall compared to s^e until $s^e - s = \Delta s^e = +1\%$, i.e., until an appreciation of the domestic currency by 1% is expected. Hence, the relation $s = s^e + i - i^*$ determines the current exchange rate depending on both interest rates and on the expected exchange rate. A practical application is presented in box 4.14, showing that a variation in the interest rate may have a magnified impact on the exchange rate in the short run depending on: (i) How long the domestic interest rate will differ from the foreign one, and (ii) what will be the impact of the monetary policy on the exchange rate in the long run. This is consistent with exchange rates generally being much more volatile than interest rates. Since 1999, for instance, euro–dollar quarterly fluctuations have typically been +/−20% on an annual basis, whereas the three-month interest-rate differential has never reached +/−3% (figure 4.12).

25. The perfect substitutability condition can be violated if the assets do not present the same risk, for example, because there is a higher probability of government default in one of the countries.

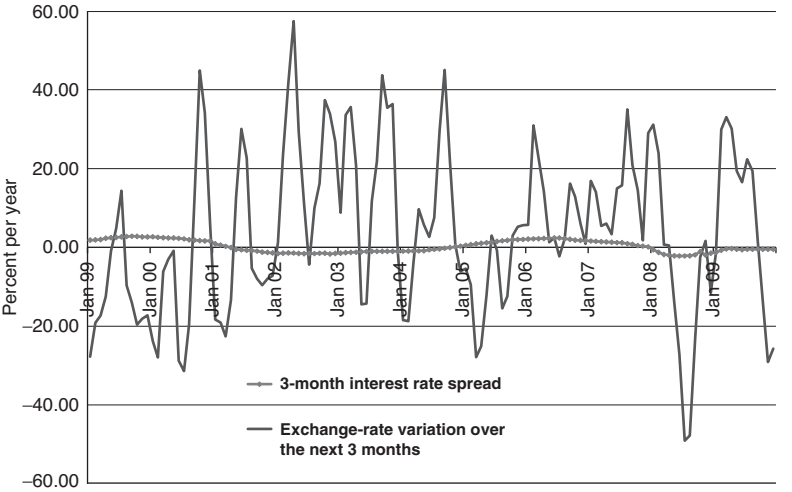


Figure 4.12 Interest-rate differentials and euro–dollar exchange rate variations, 1999–2009.
Source: European Central Bank.

Box 4.14 The Uncovered Interest Parity: A Practical Application

We start from a situation in which the uncovered interest parity (UIP) prevails: $i = i^* - \Delta s^e$. Suppose the domestic central bank decides to lower the interest rate i by one annual percentage point compared to the foreign interest rate i^* during one year. We first assume that the expected exchange rate s^e is unchanged. The domestic return falls below the foreign one, which triggers net capital outflows and a currency depreciation (s declines) until the UIP prevails again. The exchange rate depreciates by 1% on impact, and from this depreciated level investors expect an appreciation by 1% to bring the exchange rate back to its previous level, which corresponds to the unchanged expected exchange rate.

Now assume that the interest-rate differential is maintained during two years before being closed to zero (figure B4.14.1). In this case, the exchange rate depreciates by 2% in the short run. Then, it appreciates by 1% each year and returns to its initial level after two years, when the interest-rate differential is closed. It can be noted here that the initial depreciation of the exchange rate is larger than the interest-rate differential. Exchange rates are more volatile than interest rates.

Finally, assume that the fall in the domestic interest rate over two years triggers a 1% depreciation of the exchange rate in the long run, for instance because the exchange rate must compensate for higher accumulated inflation during these two years. The reasoning of the previous paragraph still applies: The exchange rate depreciates instantly until reaching a level

from which it will appreciate by 1% a year during these two years before stabilizing at its long-run level. However, the long-run level is now lower by 1%. In order for the 1% a year appreciation to bring the exchange rate to this new equilibrium, the exchange rate must depreciate by another 1% in the short run: Total short-term depreciation is now 3% instead of 2% (Figure B4.14.1).

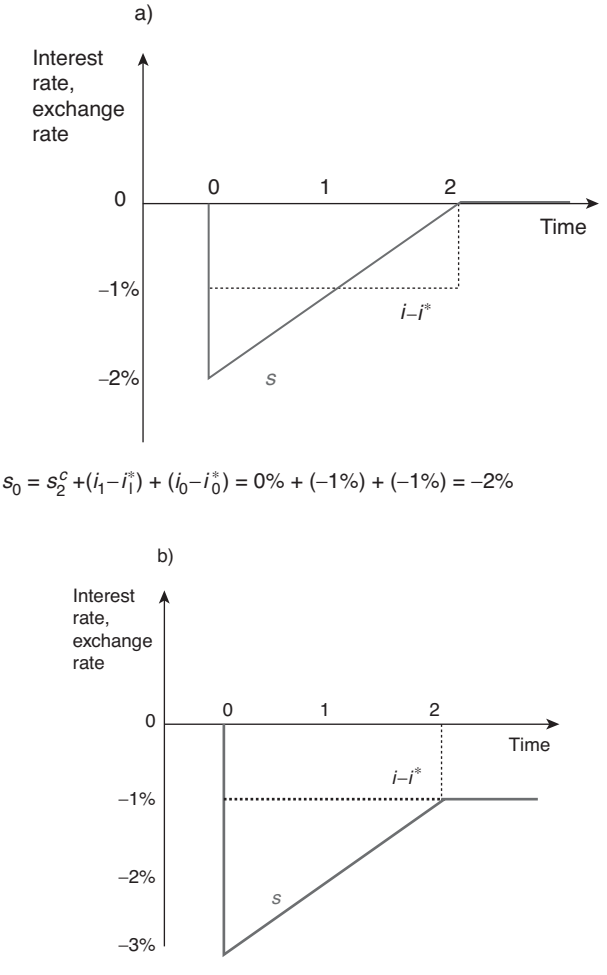


Figure B4.14.1 The impact of a one-percentage point fall in the interest rate over two years. a) Unchanged long-run exchange rate, b) depreciated long-run exchange rate.

More generally, UIP implies that the exchange rate at time t is the sum of expected interest-rate differentials from t to $t + T - 1$ and of the expected exchange rate for time $t + T$:

$$s_t = s^e_{t+T} + \sum_{\tau=0}^{T-1} \left(i_{t+\tau} - i^*_{t+\tau}\right)^e \tag{4.4}$$

Equation (4.4) implies that the exchange rate responds immediately to events that affect market expectations of future monetary policy. For example, the publication of a disappointing employment figure suggests that the central bank will reduce its interest rate or increase it later than previously expected. In reaction, the exchange rate will depreciate immediately. Hence the exchange rate at time t incorporates market expectations concerning monetary policy over the horizon.

Figure 4.13 illustrates this phenomenon. In early June 2008 the US dollar had appreciated against the euro following declarations by Ben Bernanke, the Chairman of the Federal Reserve, expressing worries about the dollar weakness. But on 5 June the President of the ECB, Jean-Claude Trichet, said that a rise in the interest rate on the euro in July was "possible." The euro appreciated immediately. The next day, the publication of worse than expected US unemployment data added to the dollar depreciation.

Although the UIP relationship cannot be tested directly because it involves unmeasured exchange-rate expectations, there are serious doubts that it applies (see box 4.15). Nevertheless, the central idea remains that, like all

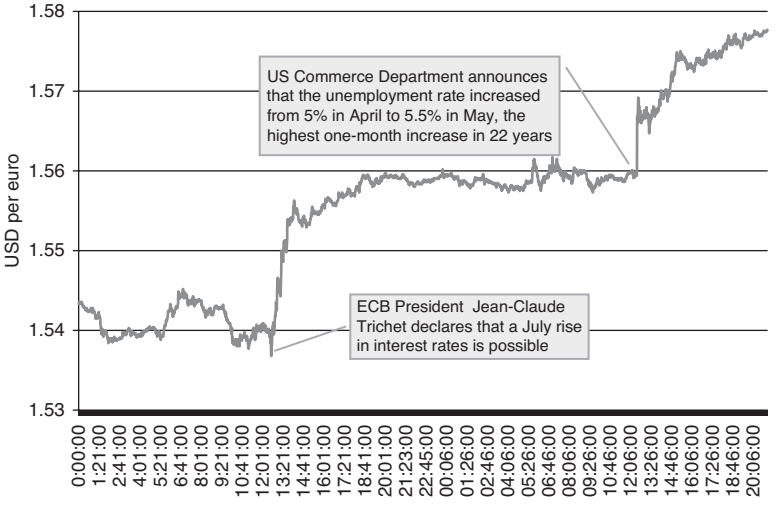


Figure 4.13 The euro–dollar exchange rate on 5 and 6 June 2008.
Source: Reuters.

asset prices, the exchange rate incorporates all relevant information about future interest rates and their determinants, and that this creates a significant volatility. From a policy standpoint, this means that disseminating information concerning future monetary policy (for instance through speeches) can move the exchange rate even in the absence of concrete monetary action. This requires discipline in handling market-sensitive information, since exchange-rate variations can have a very large impact on output and prices, depending on the structure of the economy (see chapter 5).

Box 4.15 More on the Uncovered Interest Parity

The link between the current exchange rate and expectations concerning future interest-rate differentials (equation 4.4) also holds in real terms. Denoting by $r = i - \pi^e$ the real interest rate (with $\pi^e = p^e - p$ the expected inflation rate) and by $q = s + p - p^*$, the (logarithm of the) real exchange rate, the UIP yields:

$$q = q^e + r - r^* \tag{B4.15.1}$$

hence:

$$q_t = q^e_{t+T} + \sum_{\tau=0}^{T-1} \left(r_{t+\tau} - r^*_{t+\tau}\right)^e \tag{B4.15.2}$$

Suppose we have a theory of the equilibrium real exchange rate, i.e., where the real exchange rate will settle in the medium term (such a theory is outlined in chapter 5), and assume that at time $t + T$, the real exchange rate has converged on its equilibrium level. Then, equation (B4.15.1) provides a short-term determination of the exchange rate as a function of: (i) Its equilibrium level, and (ii) expectations of future real interest rates.

Empirically, however, the UIP condition does not perform well. It is not directly observable, since the expected exchange rate is not observed, but it can be tested indirectly. Hence some assumptions need to be made to test for UIP. To see it, let us decompose the variation in the (log of the) exchange rate as follows:

$$s_{t+1} - s_t = (s_{t+1} - s^e_{t+1}) + (s^e_{t+1} - f_t) + (f_t - s_t) \tag{B4.15.3}$$

where f_t is the *forward exchange rate**, i.e., the price set at time t for purchasing or selling the domestic currency at time $t + 1$. The arbitrage condition implies that:

$$i_t = i^*_t - (f_t - s_t) \tag{B4.15.4}$$

This equality, called *covered-interest parity**, resembles the UIP condition. The difference is that there is no risk in trading-off between: (i) Investing in domestic assets (carrying an i_t return), and (ii) investing in foreign assets (remunerated at rate i_t^*) but covering the foreign-exchange risk by selling the foreign currency on the forward market (at a price f_t which is already known when the investment is decided). Because there is no risk involved in this trade (other than a country risk that is neglected under normal conditions), the covered interest rate applies in reality. This means that $i_t^* - i_t$ can substitute $f_t - s_t$ in the UIP relationship.

With this in mind, equation (B4.15.3) can be interpreted as a decomposition of the exchange-rate variation $s_{t+1} - s_t$ into: (i) A forecast error $(s_{t+1} - s_{t+1}^e)$, (*i*) a risk premium $(s_{t+1}^e - f_t)$ that measures the excess return obtained when a bet is made on the future spot rate s_{t+1}, and (iii) the interest-rate differential $(i_t^* - i_t)$. Assuming that forecast errors are nil on average (rational expectation assumption) and that the risk premium is constant over time, equation (B4.15.3) reduces to the UIP and it can be tested by estimating the following equation:

$$s_{t+1} - s_t = \alpha + \beta(f_t - s_t) + u_{t+1} \qquad \text{(B4.15.5)}$$

The uncovered interest-rate-parity condition with zero risk premium corresponds to $\alpha = 0$ and $\beta = 1$. Empirical estimates typically lead to finding $\alpha \neq 0$ and $\beta < 1$. It is even common to find $\beta < 0$ (Chinn and Meredith, 2004). This can be explained by time-varying risk premiums (Fama, 1984), nonrational exchange-rate expectations (Frankel and Froot, 1989) and/or learning processes (Gourinchas and Tornell, 2004).

c) Exchange-rate overshooting

The external channel of monetary policy is all the more powerful given that the exchange rate reacts strongly to an interest-rate change. In 1976, Rudiger Dornbusch studied this adjustment mechanism within the framework of a fully fledged macroeconomic model with *sticky prices**, i.e., where prices are rigid in the short run but flexible in the longer run. In his model, consistent with long-run money neutrality, a 1% rise in money supply leads, in the long run, to a 1% increase in prices and to a 1% depreciation in the nominal exchange rate (see box 4.16). In the short run, however, the nominal exchange rate depreciates by more than 1%: This is called *exchange-rate overshooting**. The cause is price stickiness,

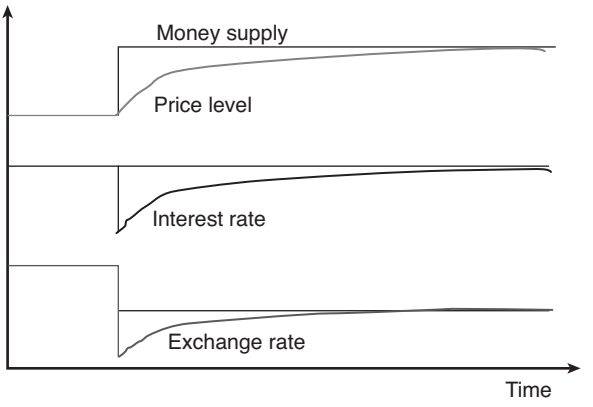


Figure 4.14 The Dornbusch overshooting model: How prices, the interest rate, and the exchange rate react to a once-and-for-all increase in money supply.

together with forward-looking expectations: Since prices do not increase in the short run, money supply rises in real terms. Hence, the interest rate falls. Along the same lines as in the previous paragraph, the exchange rate depreciates more in the short run than in the longer run. The price level then increases, which reduces the real value of the money supply: The interest rate rises back to its international level. The gradual reduction in the interest-rate differentials slows down exchange-rate appreciation. When the interest-rate differential is back to zero, the exchange rate eventually stabilizes (figure 4.14).

Box 4.16 The Sticky-Price Monetary Model of Exchange-Rate Determination (Dornbusch, 1976)

The model is based on four equations: The money-market equilibrium (B4.16.1), an aggregate-demand curve (B4.16.2), a price-adjustment relationship (B4.16.3), and the uncovered interest parity (B4.16.4):

$$m_t = p_t + \alpha y - \beta i_t \tag{B4.16.1}$$

$$d_t = \gamma y - \delta(s_t + p_t) \qquad \delta, \gamma, > 0 \tag{B4.16.2}$$

$$p_{t+1} - p_t = \theta(d_t - y) \qquad \theta > 0 \tag{B4.16.3}$$

$$i_t = i_t^* - (s_{t+1} - s_t) \tag{B4.16.4}$$

where m_t is the logarithm of money supply, d_t is the logarithm of aggregate demand, y is the logarithm of output (exogenous). Two dynamic

relationships between the exchange rate and the price level can be derived from these four equations:

$$s_{t+1} - s_t = \frac{1}{\beta}(\bar{p} - p_t) \tag{B4.16.5}$$

$$p_{t+1} - p_t = \theta\delta(\bar{s} - s_t) + \theta\delta(\bar{p} - p_t) \tag{B4.16.6}$$

where $\bar{s}$ and $\bar{p}$ represent the long-run values of the nominal exchange rate and of the price level, respectively. Consistent with long-run money neutrality, a 1% increase in money supply leads in the long run, other things equal, to a 1% increase in prices and a 1% exchange-rate depreciation:

$$\bar{p} = m - \alpha y + \beta i^* \tag{B4.16.7}$$

$$\bar{s} = -\bar{p} - \frac{1-\gamma}{\delta} y \tag{B4.16.8}$$

Equations (B4.16.7) and (B4.16.8) are represented by a horizontal line and a downward-sloping line in figure (B4.16.1). The joint dynamics of the exchange rate and of the price level are represented by the arrows based on equations (B4.16.5) and (B4.16.6). Only one locus in the graph, called the saddle path, allows for convergent dynamics, all other trajectories being divergent. We assume that the trajectory of the economy over the long run has to follow this saddle path so as to reach long-term equilibrium. We can now study the evolution of the exchange rate following a monetary shock. For instance, assume that money supply increases permanently. The horizontal line moves upward and the long-run equilibrium shifts from E_0 to E_1. In the short run, however, prices are rigid. Hence the equilibrium jumps from E_0 to E_0', which is located on the saddle path with the initial price level p_0. The nominal exchange rate s_0' is lower than both the initial level s_0 and the new, long-run level s_1: The exchange rate overshoots its long-run level. From this short-run equilibrium, the price level and the exchange rate move upward along the saddle path until the long-run equilibrium E_1 is reached.

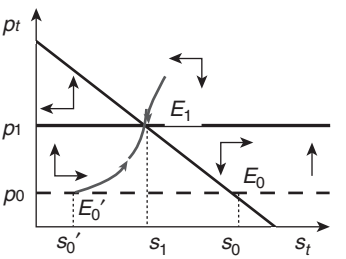


Figure B4.16.1 The dynamic adjustment of the price level and of the exchange rate following a permanent rise in money supply.

4.2.4 Financial stability

Price stability is often defined as a state in which inflation does not influence people's economic decisions.[26] Likewise, *financial stability** can be defined as a state in which the health of the financial system does not influence people's economic decisions. More precisely, it is a situation where agents can safely rely on a smooth functioning of the financial system as a whole, and are given the proper instruments and incentives to assess correctly the risk associated with particular assets they contemplate investing in. It does not need to imply public intervention to hamper market-based asset price adjustment, but requires monetary policy and regulatory policy responses to market developments.

Maintaining financial stability involves:

- An adequate infrastructure (such as the payment, clearing and settlement systems).
- A regulatory framework so that financial intermediaries are given incentive to assess risk correctly, value financial assets, set aside capital against the risky ones, and build liquidity buffers in preparation of possible market disruptions; Systematically important financial actors are properly supervised and financial centers enforce a level-playing field in regulation.
- Deposit-insurance schemes and the provision of central-bank liquidity when needed.

In exceptional circumstances, financial stability requires using public money in order to limit the risk of problems affecting particular institutions spreading to other parts of the financial system. This task usually involves the central bank, as well as supervisory authorities and ultimately the budgetary authorities, as any recapitalization of ailing financial institutions implies the injection of public money.

The reason why authorities need to preserve financial stability is that financial transactions by nature involve risks that can spill over from one institution or market segment to the market as a whole and thereby acquire a systemic character. To understand why, it is appropriate to start from the standard description of banks as performers of *financial transformation**: They receive short-run deposits which they transform into long-run lending, either directly by extending loans or indirectly by purchasing marketable securities. Such transformation involves a *credit risk** when a debtor is defaulting, a *market risk** when the value of an asset suddenly moves, and a *liquidity risk** when an asset cannot be sold to meet a reimbursement. All can result in the bank defaulting on its obligations. In anticipation of that risk, customers may withdraw their deposits, thereby precipitating the risk of a collapse. Furthermore, since banks are mutually in debt to each other (see section 4.1.1),

26. "A state in which the general price level is literally stable or the inflation rate is sufficiently low and stable, so that considerations concerning the nominal dimension of transactions cease to be a pertinent factor for economic decisions" (Papademos, 2006, p. 1).

a particular bank's default could spread to others and affect the financial system as a whole. This is the classic *bank run**, of which Walt Disney's Mary Poppins offers a canonical model (box 4.17).

Box 4.17 The Mary Poppins Model of Bank Runs

For the first time, Mr. Banks brings his children Michael and Jane to see the bank of which he is an employee, chaired by old Mr. Dawes (Senior).

Mr. Dawes: Uh, Father, these are Banks's children. They want to open an account.

Mr. Dawes (Senior): Very well, my boy, give me the money.

Michael: No, I won't! I want it to feed the birds.

Mr. Banks: Yes, sir. Now, Michael. When you deposit tuppence in a bank account [. . .] Soon you'll see [. . .] that it blooms into credit of a generous amount semi-annually. [. . .]

Mr. Dawes (Senior) & Directors: You can purchase first and second trust deeds. Think of the foreclosures! Bonds, chattels, dividends, shares. Bankruptcies. Debtor sales. Opportunities. All manner of private enterprise. Shipyards. The mercantile. Collieries. Tanneries. Corporations. Amalgamations.

(Mr. Dawes takes the two-pence away from young Michael's hand)

Michael: Give it back! Gimme back my money!

Client 1: There's something wrong. The bank won't give someone their money!

Client 2: Well, I'm going to get mine! Come along, young man! I want every penny!

Client 3: And mine, too!

Client 4: And give me mine, too!

Banker: Stop all payments. Stop all payments.

In advanced economies, bank runs were considered an historical oddity until, in September 2008, clients of Northern Rock (a British bank specialized in mortgage lending) formed queues in the street, forcing the British Treasury to issue a statement guaranteeing all deposits made at the bank. At wholesale level, US investment banks Bear Sterns (in spring 2008) and Lehman Brothers (in summer 2008) were subject to runs. Street panics would certainly have occurred in autumn 2008 after the bankruptcy of Lehman Brothers, had governments not introduced blanket deposit guarantees and had they not announced that no further systematically important financial institution (or *SIFI**) would be allowed to fail.

Bank runs and liquidity shortages are an old phenomenon but they are not easy to formalize. Models of self-fulfilling crises à la Mary Poppins can be written down easily but they fail to capture a large part of the reality. What we need to understand is why short-term financing can sometimes become

an issue of life and death. As observed by Jean Tirole (2008), the notion that a solvent institution whose spending and investment decisions are appropriate may be unable to finance them is in contradiction with standard economic theory. A solvent institution whose investments are profitable should be at any point in time able to pledge future income in exchange for immediate financing. Furthermore, liquidity is an ambiguous notion as the same term is used in two different ways: First, to say that transaction costs on an asset are low (in this sense, a stock is liquid whereas property is not) and second, to say that its price does not vary significantly in presence of macroeconomic shocks (in this sense, a cash deposit is liquid but a stock is not).

An example provided by Tirole (2008) helps us understand what is at stake. Consider an entrepreneur who engages in a risky project:

- At date 0 she/he invests 10 by drawing on her/his own resources (8) and borrowing from the capital market (2).
- At date 1, there is a 0.5 probability that a "liquidity shock" occurs;[27] in this case, the entrepreneur needs to pay 20, otherwise the project is terminated and yields no income at all.
- At date 2, revenue accrues (provided the project has not been abandoned at date 1). The total proceeds (30) are shared between investors (12) and entrepreneur (18). It is assumed that the entrepreneur cannot pledge to pay more than 12 to creditors. The interest rate is supposed to be nil, which implies that the income exactly covers the creditors' expected input ($2 + 20/2$)

The problem is that, if the shock occurs at date 1, capital markets will be unwilling to lend an additional 20 to the entrepreneur, just because the amount they are to receive at date 2 (12) does not cover this emergency loan. In order to be certain to carry out the project, the entrepreneur negotiates a credit line with a bank at date 0.This credit line ensures that the entrepreneur will obtain 20 if the shock occurs. In exchange, she/he will have to pay a fixed commitment fee of 4, even if the shock does not occur, and to transfer her/his own date-2 return (12) to the bank if the shock occurs. The fixed fee exactly covers the bank's expected loss.[28] The bank therefore provides state-contingent lending, which is a form of insurance.

Suppose now that there are many such entrepreneurs. As long as shocks are firm-specific, the private sector (entrepreneurs + banks) is able to deal with liquidity shocks. At each period, half of the entrepreneurs face liquidity shocks and borrow from banks, half do not rely on credit. Banks and entrepreneurs are profitable and no public intervention is required. If a macroeconomic shock occurs, however, the proportion of entrepreneurs facing liquidity shocks increases and the private sector is not able to generate

27. For instance, the entrepreneur is obliged to acquire a new technology that has just appeared in the market.
28. That is: $(1/2)(20 - 12)$.

the required liquidity—Tirole, in joint research with Bengt Holmström, speaks of *inside liquidity*—and the central bank has to step in.

This simple example helps us understand why central banks have a key role to play in situations of financial stress when risks that are normally uncorrelated suddenly materialize simultaneously.[29]

Similar ideas were formalized in 1983 by Douglas Diamond and Philip Dybvig to account for bank runs (box 4.18).

Box 4.18 The Canonical Model of a Bank Run (Diamond and Dybvig,1983)[a]

Consider an economy inhabited by a continuum of individuals each endowed with one unit of money in period $t = 0$, which they keep as deposit in a bank or spend on consumption. All individuals are identical *ex ante* but have a probability π of suffering a *liquidity shock* which forces them to consume in period $t = 1$ (short run) instead of waiting until period $t = 2$ (long run) when they do not suffer this shock. Their expected utility is written as:

$$U = \pi u(C_1) + \beta(1 - \pi)u(C_2)$$

where $u(C_i)$ represents the utility of consumption C_i in period i ($i = 1, 2$) and β is the discount factor ($0 < \beta < 1$). The deposits received by the bank can either be kept in liquid assets or invested in long-run assets (securities, real estate, bank loans, long-run bonds). It is assumed that only the second type of investment by the bank yields a positive return. The first one is risk-free but yields zero return. Hence the bank has a clear incentive to invest in long-run assets. The problem is that it must insure itself against the risk that some depositors withdraw their deposits at $t = 1$. Although it is possible to sell long-run assets at $t = 1$, a penalty is charged that brings their return to the negative zone. Note $R > 0$ the return on long-run assets and $r < 0$ the return on long-run assets that are sold before maturity. It is easy to show that the optimal behavior of the bank is to invest a proportion $(1 - \pi)$ of deposits in long-run assets and keep the remaining part in liquid form.

Now let us examine the behavior of individual depositors. At $t = 1$, depositors discover whether they are of the "patient" or "impatient" type, depending on whether they undergo a liquidity shock. The "patient" depositors who only care about consumption in period 2 can leave their money at the bank and consume C_2^* in $t = 2$; or they can withdraw C_1^* to keep it in liquid form until $t = 2$. It can be shown that if $\beta R \geq 1$, then $C_2^* > C_1^*$: A depositor is better off leaving his or her money in the bank. The latter therefore prepares for the withdrawal at $t = 1$ of only a fraction πC_1^* of the "patient" depositors, and invest the remainder in the illiquid asset.

29. How economic policy should address tail risks is discussed in chapter 2.

However, if a "patient" depositor who does not face a liquidity shock suspects that the others will withdraw their deposits at $t = 1$, he or she will also want to withdraw it before the bank itself becomes illiquid. Consequently, the bank needs to liquidate its long-term assets and it receives $\pi C_1^* + (1 - \pi C_1^*)r$ from this sell-off. Since this amount is lower than C_1^*, the bank goes bankrupt at $t = 1$, which confirms that the "patient, but anxious" depositor was right to withdraw at $t = 1$. Hence the panic is self-fulfilling. Now, if $\beta R < 1$, even patient individuals withdraw.

There are therefore two possible equilibria: A Pareto-efficient equilibrium where individuals trust the bank and leave their deposits until $t = 2$, and a self-fulfilling panic where individuals withdraw their deposits in $t = 1$ and where the bank goes bankrupt. Nothing in the model makes it possible to predict which equilibrium will emerge; in reality, a panic can be triggered by a rumor, or simply by the news of a first wave of withdrawals, or even by the bankruptcy of another bank.

[a]This presentation relies on Freixas and Rochet (1997).

The prevention of systemic financial crises implies that central banks step in to provide liquidity when needed. As discussed already, this involves the risk of moral hazard, i.e., imprudent bank behavior in the expectation that they will be bailed out. Fahri and Tirole (2008) provide a formalization of this risk in a simple three-periods setting where entrepreneurs choose between a safe technology (which does not involve the risk of a liquidity shock) and a cheaper risky technology (which involves the risk of liquidity shock in the second period). If they choose the risky technology and a liquidity shock occurs, entrepreneurs go bust unless the central bank lowers the interest rate, thereby bailing them out macroeconomically—at the cost of reducing consumer welfare. The model is intended to capture the effects of Alan Greenspan's policy stance, which consisted in letting bubbles develop while getting ready to support the economy when they burst.

In this setting entrepreneurs always choose the safe technology if they anticipate that the central bank will keep the interest rate constant, but they choose the risky technology if they expect a bail-out. Either the central bank credibly commits not to lower the interest rate whatever happens, or if not, the expectation of its action leads entrepreneurs to make a technology choice that will make the monetary bail-out socially optimal after the shock occurs. This is a typical time-inconsistency problem akin to the one formalized by Barro and Gordon (see above).

There are two possible answers to this dilemma. One is to exclude the possibility of a bail-out. In practice, this is hardly credible, because agents know that the central bank will be forced to act if the shock is severe enough (this was confirmed in 2008), unless the size of the banks is strictly capped so that none of them are systematically important (see chapter 8). The other

answer is to complement monetary policy with a regulatory instrument which ensures that risky investment is capped at a certain level. In a nutshell, this is the role of financial supervision. A related proposal consists in banning deposit-taking banks from engaging in risky activities such as proprietary trading of financial assets, as advocated by former Fed Chairman Paul Volcker and endorsed by President Obama in January 2010 (see chapter 8).

4.3 Policies

4.3.1 Institutions

Monetary policy is everywhere carried out by the central banks.[30] However, this does not imply that they have full responsibility for decisions. Until the late 1980s, only a few countries—most notably Germany—had a fully independent central bank.

a) The move to central bank independence

Institutional change started in the 1990s and accelerated in the 1990s and the 2000s as an increasing number of countries granted full *independence** to their monetary institution, as shown in figure 4.15.

This move resulted first and foremost from the better ability of independent central banks to cope with the inflationary pressures of the previous decades: In the worldwide competition between two institutional models, central bank independence had won. However, it also built on the theoretical rethinking of the previous decades. In particular, rational expectations theorists Robert Lucas and Robert Barro (see chapter 2) had produced models with rational expectations where no trade-off between inflation and unemployment was possible, even in the short run. A particularly forceful illustration was the Barro–Gordon model presented in box 4.10, where attempts by the central bank to exploit a short-run trade-off are anticipated by the private agents, which results in an inefficiently high inflation and no employment gains. This suggested, first, that there was no cost in specializing the central bank in fighting inflation, with no role for output stabilization and, second, that there were gains in making it independent from government. Hence, monetary policy could be delegated to an independent agency with a clear mandate to guarantee price stability.

Another argument in favor of central bank independence is that the government itself may have some interest in engineering higher inflation. The central bank manufactures banknotes at low cost and issues them at their nominal value. Notes are thus akin to an interest-free loan extended

30. That is, insofar as the regulation of broad money is concerned. Some stores of value and means of exchange are not controlled by the central bank and may at times gain quantitative importance, such as consumer credit or frequent-flyer mileage.

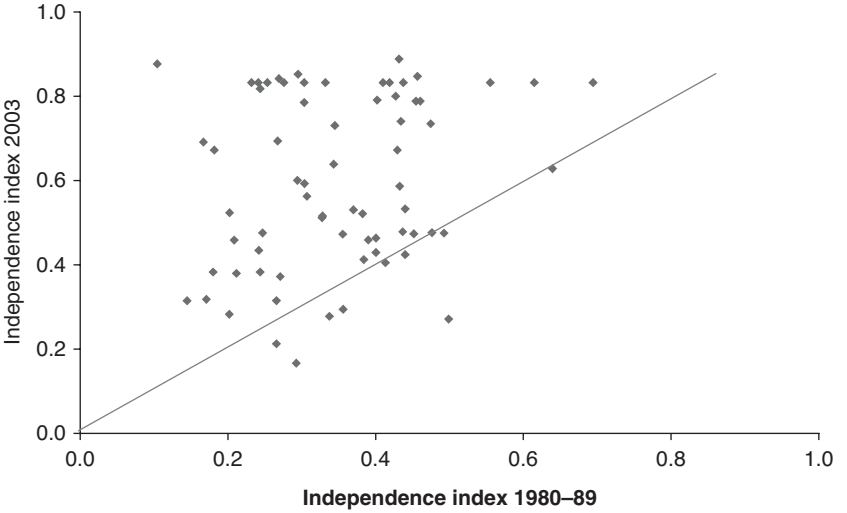


Figure 4.15 Central bank independence in the 1980s and the 2000s.
Source: Crow and Meade (2008), based on Cukierman et al. (1992).
Note: Central bank independence index based on four components, relating to, respectively, appointment procedures for the head of the central bank, the resolution of conflict between the central bank and the executive branch of government, the use of an explicit policy target, and rules limiting lending to government.

by economic agents to the central bank. The income from money creation is called *seigniorage** and accrues to the central bank's profit, which is partly or entirely given back to the treasury as a dividend. If the central bank is not independent, the government can be tempted to maximize this revenue.

Seigniorage thus is a tax levied on money holders and sometimes dubbed an *inflationary tax**. However, it has become marginal as a fiscal revenue in most countries, due to the low rate of inflation (box 4.19). In countries where inflation is low, seignoriage revenue is material only to the issuing of coins. The Bank of Canada has estimated that each 1$ coin generates 88 cents in seigniorage for the government while a $20 bill, which lasts about three years, generates an annual net revenue of only 96 cents for the Bank.

Box 4.19 The Value of Seigniorage

There are two standard measures of seigniorage. The first one is the increase in base money ΔM_0 over a given period. The second measure is the opportunity cost of holding base money, i.e., iM_0, where i denotes the short-run, nominal interest rate over the same period. As detailed below, the two measures are equivalent, provided the velocity of money is constant and the real interest rate is equal to the real growth rate.

Remember the quantity equation of money of box 4.4, $PY = M_0V$ where V is the velocity of money. In the long run, if real output growth (as determined by the production function) is g and the velocity of money is constant, then the growth rate of base money $\Delta M_0/M_0$ is equal to the growth rate of prices ($\pi = \Delta P/P$) plus the growth rate of real output g. We therefore have:

$$\Delta M_0/M_0 = \pi + g \qquad \text{(B4.19.1)}$$

Now, the nominal interest rate i is the sum of the real interest rate r and the inflation rate: $i = r + \pi$. If the real interest rate r is equal to the real growth rate g (golden rule of capital accumulation), then equation (B4.19.1) can be rewritten:

$$\Delta M_0/M_0 = i \qquad \text{(B4.19.2)}$$

or, equivalently:

$$\Delta M_0 = iM_0 \qquad \text{(B4.19.3)}$$

The two measures of seigniorage are then equivalent.

In the euro area, currency in circulation was €600bn in 2009, i.e., 6.8% of annual GDP; with a short-run interest rate of 4%, this yielded a seigniorage revenue of only 0.27% of GDP. In the US, the monetary base was $820bn, i.e., 5.9% of GDP in 2007; with a 5% interest rate, the seigniorage amounted to 0.30% of GDP. It is clear from Equation (B4.19.1) that seigniorage is higher in fast-growing, high-inflation countries.

However, seigniorage also depends on the velocity of money and it may not be entirely transferred to the government, depending on the central bank's dividend policy. In the euro area, since 2003, the monetary income raised by the Eurosystem has been pooled and distributed to national central banks, not according to their contribution to the euro area monetary base but according to their weight in the ECB capital, itself based on each country's GDP and population.

There are other components to the inflationary tax. To the extent that public debt is not indexed to inflation, an unanticipated inflationary shock automatically reduces the public debt burden (debt-to-nominal GDP ratio). This amounts to a transfer from asset holders to the government. Also, there is an inflationary tax levied on corporations. Corporate taxes are computed on the basis of nominal (rather than real) income, which is overstated in times of high inflation because capital depreciation (which is deducted from the tax base) is based on historical, rather than current cost.

Even though most economists today would not go so far as to deny any impact of monetary policy on real variables in the short term, it is widely recognized that well-designed central bank independence has the benefit

of enhancing monetary policy credibility and enforcing price stability at a low cost.

There are, however, notable discrepancies among independent central banks in the design of their statutes and mandates. Even leaving aside the Fed (which has a broad mandate as indicated in table 4.1) and the Bank of Japan (which is not independent but "autonomous" according to the 1997 Bank of Japan Act), the ECB and the Bank of England differ in several respects. Price stability is legally the main objective for both, and both enjoy full independence in the conduct of monetary policy, but, while the ECB itself sets the quantitative inflation objective, the Bank of England merely implements an objective decided by the government. The 1998 Bank of England Act indicates that the Chancellor of the Exchequer may specify in writing to the central bank "what price stability is to be taken to consist of." And in fact the Chancellor does, but the obligation of doing it in writing prevents it from pushing for a higher inflation without telling the public. This provision removes the temptation of surprise inflation and in fact represents a constraint for the government, which cannot criticize the bank for merely implementing a policy it has defined.

b) Statutes and mandates

To establish its reputation, the central bank must do what it says and say what it does. Hence, reputation basically relies on (i) track record, and (ii) communication. Nevertheless, theoreticians have suggested a number of schemes that can help central banks to build up their credibility (see section 4.2).

- *Independence.* In a majority of the countries, the independence of the central bank is guaranteed by law. This is a limited guarantee. In the US, for example, the Constitution gives Congress authority over the value of money. Congress can thus change the Fed's mandate by simple majority. The institution's independence is actually protected by various counter-powers, especially financial markets and the press, as well as by its own reputation. From a legal point of view, the ECB is much more independent, since its independence is part of the EU Treaty and has therefore a supranational value: Only a unanimous decision by the EU Member States, followed by ratification by their parliaments or by popular vote, would make it possible to repeal this independence. The EU Treaty (article 130 TFEU) stipulates that: "When exercising the powers and carrying out the tasks and duties conferred upon them by this Treaty and the Statute of the ESCB, neither the ECB, nor a national central bank, nor any member of their decision-making bodies shall seek or take instructions from Community institutions or bodies, from any government of a Member State or from any other body. The Community institutions and bodies and the governments of the Member States undertake to respect this principle and not to seek to

influence the members of the decision-making bodies of the ECB or of the national central banks in the performance of their tasks."

- *Commitment.* The solution consisting in tying one's hands was adopted in the 1980s by several European countries, including France, The Netherlands, and Belgium, which anchored their currencies to the German mark within the framework of the European monetary system. A similar solution was used by Argentina in the 1990s through the adoption of a "convertibility law," which anchored the currency to the US dollar (see chapter 5), but this ended in catastrophe. More recently, several transition countries adopted a similar approach and some of them even dropped their monetary sovereignty altogether by adopting either the dollar (Ecuador, El Salvador . . .) or the euro (Kosovo, Timor-Leste) as their domestic currency (see chapter 5).
- *Conservative central bankers.* Kenneth Rogoff (1985) suggested eliminating the inflation bias by appointing a "conservative" central banker, i.e., a governor whose willingness to trade off output stabilization for price stability would be higher than social preferences. This proposal has not been legally implemented but is often relied on de facto: Central bank governors are generally selected for their anti-inflationary convictions.
- *Incentive contracts.* Incentive contracts à la Carl Walsh (1995) are not frequent. In most countries, central bank governors can be dismissed only for crime, misdemeanor, or offense to the bank's reputation. In New Zealand, the Reserve Bank Act of 1989 stipulates that the central bank governor can be dismissed in the event of inadequate pusuit of the objectives (box 4.19). A less radical solution, adopted in the UK, is to compel the governor to justify himself publicly when the 2% inflation target (see above) is missed. Such an open letter was sent for the first time on 17 April 2007 by Governor Mervyn King to Chancellor Gordon Brown, explaining why British inflation in the previous 12 months was 3.1%, more than one percentage point above the Bank of England target.[31]

Box 4.20 The Incentive Scheme of the Reserve Bank of New Zealand[a]

In the 1989 Act, the Bank was given the ability to adjust the instruments of monetary policy without any routine political involvement. As part of this shift, responsibility for the exercise of the Bank's powers and the conduct of its functions was vested explicitly in the Governor. The explicit intention, in making this change, was to provide a clear focus for accountability.

31. The Governor subsequently sent similar letters every three months. The letters are available on the Bank of England's Web site. For a description of the Bank of England's remit, see HM Treasury (2001).

The Board's new role focused on two dimensions: Advising the Minister on the appointment (and reappointment) of a Governor, and monitoring and providing advice on the Governor's performance.

The Governor and Board members are both appointed by the Minister of Finance. However, whereas the parliamentary term is three years, Board members are appointed for staggered five-year terms. The Minister cannot appoint as Governor someone whom the Board has not recommended.

Before appointing a person as Governor, the Minister is required to "fix, in agreement with that person, policy targets for the carrying out by the Bank of its primary function during that person's term of office."

Section 15 of the Act requires that the Bank deliver to the Minister and publish at least every six months a monetary policy statement. In this document the Bank is required, inter alia, to review and assess recent monetary policy and to articulate "the policies and means by which the Bank intends to achieve the policy targets."

The Minister may seek the removal of the Governor (or the Board may recommend that the Minister do so) if he is dissatisfied on any of several counts. These include, inter alia, the following which bear directly on monetary policy:

- That the Bank is not adequately carrying out its functions (the primary function being monetary policy); or
- That the performance of the Governor in ensuring the Bank achieves the policy targets has been inadequate; or
- That a *Monetary Policy Statement* is inconsistent in a material respect with the Bank's primary function, or with any policy target fixed in the Policy Targets Agreement.

Note that the Act does not allow the Governor to be dismissed simply for failing to meet the policy targets. The criteria in the Act refer explicitly to the performance of the Bank and the Governor *in pursuit of* those targets.

[a]Source: Michael Reddell, "Monetary policy accountability and monitoring," Reserve Bank of New Zealand Web site.

Whether or not some of these schemes are in place, the central bank fundamentally derives its legitimacy from its ability to fulfill its targets and from its *accountability**, i.e., its exposure to external scrutiny and its answerability vis-à-vis its principal. In most countries, central bankers regularly appear before parliamentary committees to explain their policy and respond to criticism. In the US, pursuant to the Humphrey–Hawkins Act, the Federal Reserve reports to Congress, annually on its activity and twice-yearly on monetary policy. In Europe, the President of the ECB testifies quarterly in front of the Economic and Monetary Committee of the European Parliament,

but the Parliament can only criticize, not influence its policy. Similarly, all central banks have to report in detail on their analysis of the economy and of inflation.

However, central banks differ concerning transparency. The Bank of England publishes the minutes of the meetings of its monetary-policy committee, including individual votes, while the Fed publishes an anonymous report. The ECB does not publish the minutes of the Governing Council's meetings but the President gives a press conference immediately after the meetings, including a questions and answers session.

c) Monetary policy committees

A less-spectacular evolution of central banking in the 1990s and 2000s has been the move toward collegial decision-making through *monetary committees**. In the euro area, for instance, monetary decisions are taken by the *Governing Council* that includes the six members of the Executive Board (who are working permanently in Frankfurt) as well as all national governors of the euro area. In the US, the *Federal Open Market Committee* (*FOMC*)* brings together the seven members of the Board of Governors of the Federal Reserve System, the President of the Federal Reserve Bank of New York, and four of the eleven Reserve Bank Presidents (who rotate on an annual basis). Since the mid-1990s, a number of countries (e.g., the UK, Japan, Sweden, Norway, Switzerland, Brazil) have switched from individual policymaking (the governor alone) to monetary committees.[32] Such evolution can be viewed as a by-product of central bank independence. As Blinder (2007) puts it:

> When the central bank was just following orders communicated by the government, there was not much reason to have a committee on the other end of the phone.
>
> A. Blinder (2007), p. 107

The main justification for the use of monetary committees is that, due to a broader access to information, a committee has a higher probability of taking the right decisions than has a single governor. This is what the Condorcet jury theorem (see chapter 2) states under the following assumptions: (i) Costless information; (ii) common preferences across committee members; (iii) sincere voting; (iv) no pre-voting communication. Each of these assumptions can be debated, however. In particular, it has been highlighted that a large committee may not perform better than a single governor, because each committee member has a small probability of being pivotal (i.e., able to move the majority), which reduces the incentive to seek information, or due to higher costs when aggregating members' points of view.

32. A survey by Pollard (2004) indicates that 79 out of 88 central banks surveyed make policy decisions through a committee.

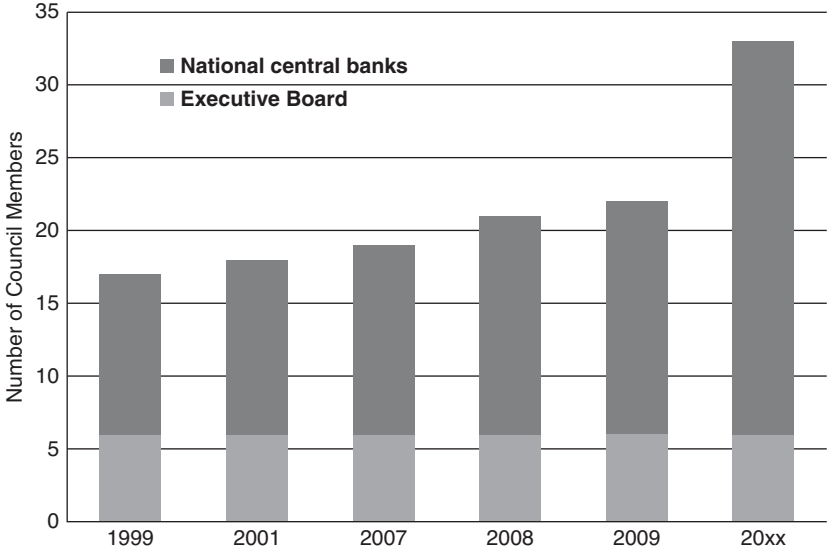


Figure 4.16 Size of the eurosystem's governing council.
Source: European Central Bank.

Social experiments tend to confirm that committees do make better decisions than single individuals, although there is no consensus on the optimal size of the committees.[33]

Another question is the composition of monetary committees, and in particular the representation of regional or national governors. If all governors have exactly the same preferences and if all have a probability $p \geq 0.5$ to take the right decision, having a wide representation of regions or countries can only be beneficial as it brings in additional information. A large committee also allows each committee member to come back home and explain the decisions to the local press in its own language. However, there is a risk that regional governors have a regional bias, which may lead to ill-designed monetary decisions.

This risk has been especially highlighted in the euro area case, where enlargement to new members will by construction increase the weight of national governors in monetary decisions. With 16 member countries in the euro area, national governors represent $16/22 = 73\%$ of the Governing Council (see box 4.1). This proportion could theoretically reach $27/33 = 82\%$ if the euro area were to be enlarged to the 27 member states of the EU (see figure 4.16).

In comparison, US regional governors represent $4/12 = 33\%$ of the Federal Open Market Committee (FOMC). To reduce the risk of monetary decisions being twisted by a majority of national governors against the executive board

33. See Blinder and Morgan (2005), Siebert (2006).

as well as by large euro area members, a system of rotation can be applied to national governors.[34] The influence of national governors in the EU will nevertheless remain higher than that of US regional governors.

Committee-based decisions raise specific questions concerning communication and transparency. Obviously, there is a risk that disorganized communication by different committee members confuses the markets. In addition, transparency can be reduced by the use of monetary committees unless the minutes of the meetings are published. In practice, however, monetary committees are not systematically less nor more transparent than single policymakers.[35]

d) Central bank–Treasury relationship

A separate issue is whether the central bank coordinates with the government, or at least maintains a dialogue with it on policy analyses and priorities. We discussed in section 4.2 of this chapter the interaction between monetary and fiscal policy, both in the short run and in the long run.

In his memoirs, Robert Rubin, the former Treasury Secretary of Bill Clinton, states that "the Treasury and Fed services work narrowly together on a series of subjects, in spite of what was historically a certain degree of institutional competition." When he was Secretary, he had breakfast "at least once a week" with Alan Greenspan. "Our discussions were very varied [. . .], we discussed quite simply the economic situation in the United States and in the world. These meetings were somewhere between an academic seminar and an operational meeting, the whole intersected with some gossips." (Rubin and Weisberg, 2003). In the UK, a Treasury representative attends all meetings of the central bank's Monetary Policy Committee where his role is especially to explain "the Chancellor's thinking on how fiscal policy will be operated" (O'Donnell, 2001).

In the euro area, the relations between the political sphere and the central bank are less intimate. The Commissioner responsible for economic affairs and the President of the euro area finance-minister meeting, the Eurogroup, can attend the Board of Governors of the ECB. However, they do not attend the dinner the night before in which policy decisions are usually discussed among central bankers. In turn, the President of the ECB generally attends the meetings of the Eurogroup, where finance ministers gather once a month to discuss the economic situation and policy priorities. There is nothing like informal, weekly breakfasts. Eurogroup president Jean-Claude Juncker called in spring 2006 for opportunities for closer dialogue but he was rebuked by ECB President Trichet since the ECB has always claimed that "*ex ante*

34. See the decision of the European Council on 21 March 2003. The rotation system can be put in place since the euro area has reached 16 countries (i.e., since January 2009) but may be postponed until there are 19 countries in it.

35. See Blinder (2007).

co-ordination" with political authorities would infringe its independence. The Federal Reserve is generally more sympathetic than the ECB to the pro-coordination view:

> A central bank, either alone or in cooperation with other parts of the government, retains considerable power to expand aggregate demand and economic activity.
>
> Ben Bernanke (2002)

The UK view goes one step further as it assigns to the government a leadership role. However, other central banks generally refute this view and reject co-ordination for the sake of their independence. In the case of the euro area, there is nothing constituting joint management of aggregate demand:

> We believe that our independence is absolutely decisive. It is enshrined in the Treaty. . . . As regards the dialogue between the Central Bank, the Commission and the Council: We in the Governing Council of the ECB organize a dialogue every fortnight to which the President of the euro group and the Commissioner concerned are invited. The Vice-President and I have the privilege of being invited by the euro group every month. . . . There is a physical dialogue which fully complies with the principle of independence.
>
> Jean-Claude Trichet (2005a)

In such a view, it is enough for each policy player to keep his or her own house in order by sticking to stated policy principles. This provides sufficiently clear information for the other players to take their own decisions. Any coordination beyond a mere exchange of information would risk blurring the objectives followed by policymakers without clear benefits in exchange. This vision of coordination by default, long advocated by Germany, is now shared by the majority of the European member states. The Maastricht Treaty, which does not envisage explicit coordination between monetary and fiscal authorities, but only exchanges of information, is consistent with this view.

e) Financial supervision

Central banks are involved in financial-crisis prevention and resolution. Prevention involves the supervision of financial institutions (banks, insurance, pension funds, etc.), which must not be confused with the supervision of financial markets. In the US, until the financial crisis of 2007–09, these two responsibilities were entrusted to many institutions: As many as five federal institutions were responsible for banking supervision (the Federal Reserve; the Federal Deposit Insurance Coroporation, the Office of Thrift Supervision, the Office of the Comptroller of the Currency, and the *Securities and Exchange Commission* (*SEC*)*), while insurance supervision was handled by State supervisors, and financial market supervision was divided between the SEC and the Commodity Futures Trading Commission (CFTC). Some countries

such as Spain and The Netherlands assign banking supervision to the central bank. In a number of countries, however, all financial supervisory responsibilities are centralized within a single institution, called the Financial Services Authority (FSA) in the UK and in Japan, and the *Bundesanstalt für Finanzdienstleistungsaufsicht* (BaFin) in Germany, etc. Even though there is no single model, banking supervisors coordinate through the *Basel Committee for Banking Supervision (BCBS)** supported by the Bank for International Settlements, to share information and draw general rules that are then written into national laws. Financial supervisors, central banks, and treasuries coordinate through the *Financial Stability Board** (FSB). In 2009, the FSB was tasked by G20 leaders with the drafting of new regulations to strengthen the supervisory system on a coordinated basis. Memberships of the BCBS and FSB have been broadened in 2009 to include major emerging market economies.

Finally, at the EU level, there remain 27 banking supervisors but the crisis has prompted reforms of the supervisory architecture both in Europe and in the US, see chapter 8.

The main rule as regards banks' capital adequacy is the *capital adequacy ratio**, initially called *Cooke ratio**.[36] This stipulates that the ratio of the bank's so-called tier-one capital (i.e., shareholders' equity plus retained earnings) to the amount of the loans granted, weighed up according to the level of risk of the counterparts, needs to be at least 4% (8% when tier-two capital, i.e., preferred shares and a fraction of subordinated debt, is added). This ratio has been revised by the so-called *Basel II accord** of 2004 and further revision was underway in 2010. The associated capital adequacy ratio distinguishes between operational risk, market risk, and credit risk, allows for finer risk discrimination based in particular on the ratings produced by credit rating agencies, and allows banks to use their own internal models to assess risks.

These new rules were intended to give more responsibility to bank managers by decentralizing risk control. Like the previous ones, they have been criticized for being procyclical: In an economic downturn, the capital ratio deteriorates, leading banks to reduce bank lending, which in turn reinforces the economic slowdown. *Pro-cyclicality** is best defined by the remark often attributed to Mark Twain that "a banker is a fellow who lends you his umbrella when the sun is shining, but wants it back the minute it begins to rain". Together with liquidity, it was at the core of the post-crisis discussions and proposals on the strengthening of the global financial system (see Chapter 8).

4.3.2 Key policy choices

a) What inflation objective?

What is the appropriate level of inflation? According to Edmund Phelps (1973), the optimal inflation rate results from a trade-off between the

36. Peter Cooke was the Chairman of the Basel committee when this ratio was defined, in 1988.

distortions stemming from the inflationary tax and those arising from other taxes. However, because the inflation tax is paid by all sectors, including the underground economy (where transactions are often carried out in cash), it can be optimal to maintain moderate inflation in a country where underground activities are quite developed.

However, this argument is secondary to developed economies. For them, there is no general recommendation except that, like the porridge of *Goldilocks*, which needs be neither too hot nor too cold, inflation has to be just right: Neither too high, nor too low. This raises statistical (box 4.21) but also theoretical difficulties.[37]

Box 4.21 The Pitfalls of Measuring Consumer Prices

Measuring inflation is difficult. Macroeconomics deals with the aggregate price level, but in the real world, only specific prices are observed. Surveys provide prices for individual items, but products change with increasing frequency. From one period to the next, new products appear and quality improvements are brought to existing products. For instance, when a computer model is replaced for the same price by another one which benefits from increased speed and memory, it is reasonable to conclude that computer prices are *falling*. But by how much? *Hedonic price** methodologies help, correcting for quality improvements by decomposing each product into a bundle of services offered by the product. However, they are applied to a limited number of goods. In addition, there are other sources of bias in the measurement of inflation, such as the development of new distribution channels or shopping habits.

For these reasons it is generally considered that official figures tend to over-estimate inflation. This observation is important for monetary policy since it means that central banks seeking very low inflation rates will face the risk of triggering deflation, i.e., a decrease in the aggregate price level.

In 1996, at the request of the US government, an independent commission chaired by Michael Boskin (Stanford University) studied the construction of the US Consumer Price Index (CPI) and concluded that its growth rate was over-estimated by 0.8 to 1.6% a year, an overall bias often dubbed the *Boskin effect** (Boskin et al., 1996). Four different biases were identified:

- *Product substitution*: The composition of the basket was not revised sufficiently often to follow quick substitutions by consumers, for instance in reaction to relative price changes;
- *Changes in shopping habits*: For instance, the fast development of e-shopping was not accounted for;

37. On the optimal level of inflation, see also Wyplosz (2001).

- *Quality improvements*: The measurement of prices did not account for quality improvements, for example when a car is sold at the same price but with more accessories;
- The appearance of *new products*.

The US government thereafter introduced a yearly revision of the consumption basket. Consistently, a second study performed by the Federal Reserve in 2001 found a measurement bias of only approximately 0.6% (Lebow and Rudd, 2001). Similar measurement biases have been estimated in other countries. For instance, the bias could lie between 0.35 and 0.8% per year in the UK (Cunningham, 1996) and be 0.9% in Japan (Shiratsuka, 1999).

Beyond these measurement difficulties, consumers' perception of inflation can differ significantly from the statisticians' measurements. For instance, consumers pay more attention to the prices of those goods they need to purchase frequently (food, gas) than to those they purchase infrequently (insurance, durable goods).

In a full-information, rational-expectations setting, economic agents would be almost completely indifferent to the level of inflation—but for the cost of withdrawing cash more frequently from the nearest ATM in order to avoid holding too many noninterest-bearing banknotes. However, high inflation erodes all nonindexed incomes, be they wages, pensions, or fixed-income revenue and is therefore likely to penalize poor people disproportionately. In a high-inflation context, agents tend to protect themselves from purchasing power loss through entering into indexed contracts (wages, for example, can be indexed to the evolution of prices), invoicing in foreign currencies, holding foreign assets, etc. This behavior was evident in Germany in the 1920s and in numerous developing countries in the second half of the twentieth century. Inflation therefore generates real rigidities and distortions, which in turn make it more difficult to reduce inflation.

Another argument against inflation is that a high inflation *level* generally comes hand-in-hand with inflation *variability*, which creates noise and distorts economic decisions, since agents may confuse variations in inflation for variations in relative prices. Also, inflation variability results in expectation errors. This can hamper investment, and hence long-term GDP growth. Empirically, Bruno and Easterly (1996) found that GDP growth is reduced when inflation is higher than 20–40% a year. Robert Barro (1997) finds that other things being equal, a rise in the inflation rate by 10 percentage points eventually reduces annual growth by about 0.3 percentage points (while noting that this result does not apply to low inflation rates, e.g., a move from 2% to 3% inflation).

Inflation can also be too low, for three reasons that have to do with information asymmetry or nominal rigidities.

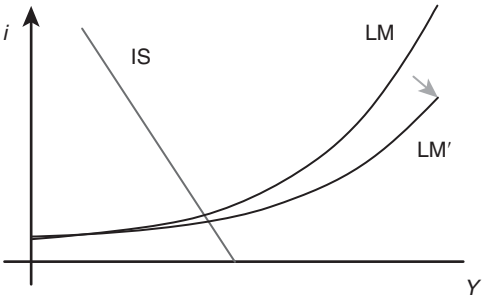


Figure 4.17 The liquidity trap.

First, near-zero inflation can result in a *liquidity trap**. At very low levels of inflation and nominal interest rates, monetary policy loses traction. This is because asset holders no longer choose between holding money and interest-bearing securities when the latter's yields are too low to compensate for lower liquidity and higher risk. Asset holders only demand money (or real assets such as stocks and real estate). In the IS–LM model, this translates into an LM curve (money market equilibrium) which is shaped as in figure 4.17: When the nominal interest rate i is very low, a monetary expansion (shift of the LM curve to the right) has no impact on output Y.

Second, low inflation can be detrimental because, provided nominal money can be stored at no cost, the nominal interest rate cannot be negative—an observation first made by Irving Fischer in the 1930s.[38] Economists thus speak of a *zero bound** on interest rate policy (see chapter 8 for an application to 2009). Hence, the real interest rate cannot fall below the opposite of the inflation rate. The more negative inflation is, the higher the real interest rate is, whereas a decline in the real interest rate would be needed instead.

The liquidity trap and the zero bound on nominal interest rates were considered as mere theoretical and historical curiosities until Japan experienced the vicious cycle of deflation in the late 1990s. In the aftermath of a collapse in asset prices, consumer-price inflation dropped below zero in 1999 and remained in negative territory for seven consecutive years. The Bank of Japan first brought interest rates to zero, but without effect on the economy. In a famous article, Paul Krugman (2000) explained why Japan needed "a credible commitment to expand not only the current but also *future* money supplies, which therefore raises expected future prices—or, equivalently, a credible commitment to future inflation" that would lower the real (*ex ante*) interest rate. In fact, the Bank of Japan had to switch to quantitative monetary

38. Negative short-term interest rates would mean the bank charging customers a proportional fee for holding a deposit. This has actually happened in rare instances—for example in Japan in the 1990s, and in Sweden in the 2000s—but those fees cannot be significant, since customers can alternatively keep cash in a safe at home.

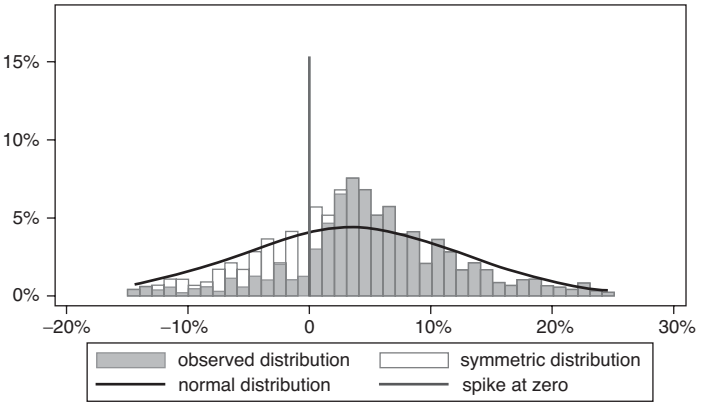


Figure 4.18 Distribution of wage changes in the US in 1988.
Source: Dickens et al. (2007).
Note: The figure gives the distribution of individual wage changes and compares it to a normal distribution. The distribution is skewed towards positive changes and presents a spike at zero—consistent with existence of nominal rigidities.

expansion in March 2001. It decided to increase the monetary base until inflation reached positive territory—a kind of money-targeting rule.[39]

The Japanese experience prompted contingency planning in the US where a quantitative strategy was also envisaged in 2002 by the Fed. Its governors felt the need to make clear that:

> By increasing the number of US dollars in circulation, or even by credibly threatening to do so, the US government can also reduce the value of a dollar in terms of goods and services, which is equivalent to raising the prices in dollars of those goods and services.... The Fed could enforce these interest-rate ceilings [on longer-maturity Treasury debt] by committing to make unlimited purchases of securities up to two years from maturity at prices consistent with the targeted yields.
>
> Ben Bernanke (2002)

Third, inflation can be too low even in the absence of deflation risks. This is because wages tend to be rigid downward: Labor contracts rarely consider revising wages downward (figure 4.18). In this context, changes in relative prices made necessary by shocks or trend changes in equilibrium prices are easier in a moderate inflation context: Inflation is like oil in the wheels of nominal adjustments. For instance, a fall in real wages can be engineered more easily by keeping nominal wages constant while prices are rising. The downward rigidity of wages is documented in the US by Akerlof et al. (1996, 2000) who argue that at near-zero inflation rates the Phillips curve, which

39. On the liquidity trap in Japan, see Krugman (1998, 2000), Ueda (2000), and Svensson (2003).

relates unemployment to the inflation rate, is not strictly vertical even in the long run: If nominal wages are rigid downward, some inflation allows reducing the unemployment rate. In view of these arguments, Akerlof et al. (2000) advise an inflation "band" of 1.5% to 4% per year in developed countries.

On the whole, a broad agreement exists today that central banks should adopt a low, but positive inflation objective. There are discussions on what the numerical objective should be, but within a narrow range: Say, between 1/1.5% to 3/3.5% for developed economies. Blanchard et al. (2010) elicited an outcry among central bankers when they suggested raising the target to 4% rather than 2% to create some more policy space for interest-rate policy.

b) What monetary strategy?

A monetary strategy is a policy framework that relates instruments to objectives. A major difficulty in this respect is that the instrument (the interest rate) only affects the final objective (inflation) with "long and variable lags," to quote an expression first introduced by Milton Friedman. The central banker must act as a sailor whose steering changes the course of the ship only gradually and not in a perfectly predictable way, and who permanently runs the risk of giving too much or too little impulse. The questions are what kind of outside information monetary policy should react to in order to best keep up with its course, and what information it should neglect; and how it should formulate and communicate its strategy in a way that helps agents form expectations about the future course of monetary policy.

Since the 1960s, debates about monetary strategy have never faded away. There are not many discussions between the advocates of quantitative (i.e., money-supply) strategies and those of interest-rate strategies anymore, but the debate has moved on.

Let us assume that the central bank intends to keep inflation at a certain level, say 2%. How should it set the interest rate? Lars Svensson (1999, 2001) proposes distinguishing between three types of rules:

- *Instrument rules**, which express the instrument(s) as a prescribed function of predetermined or forward-looking variables such as inflation or the output gap. The Taylor rule, which determines the interest rate as a function of inflation and the output gap (see section 4.1.2), is an example of such a rule. It is often used as a benchmark to assess the tightness of monetary policy, but in practice no central bank follows a mechanistic rule of this kind.
- *Targeting rules**, where the central bank aims at minimizing a loss function, such as a weighted average of square deviations of objective variables from target values. Such a rule specifies what the objectives of the central bank are and which trade-offs it can enter into, if any. The inflation-targeting strategies currently implemented by several central banks are examples of such rules. They have the advantage of

introducing transparency about the objective and the strategy while giving the central bank more discretion than a pure instrument rule.

- *Intermediate-targeting rules*, according to which the central bank attempts to control an intermediate target variable that is highly correlated with the goal but easier to observe and to control than the goal. This was the role assigned to monetary aggregates in the strategies of the 1970s, before the relationship between these aggregates and inflation broke down.

Svensson further distinguishes between explicit rules, when the variables entering the reaction function can be observed—as in a Taylor rule—and implicit rules, when some variables entering the reaction function are anticipated and not directly observable.

The dilemma to address is between relevant, but not directly observable, variables (such as inflation and production 18 months ahead) and directly observable, but less relevant, variables (such as current inflation and the current output gap).

*Money targeting** rules of the third type above were widespread from the 1960s to the 1980s, and Germany officially targeted money aggregates until 1999, though in a loose way. However, since financial deregulation and technological change have unleashed a wave of financial innovations in the 1980s, the relationship between money and inflation started to crumble and became looser and looser. Both the Federal Reserve and the Bank of England were quick to play down the importance of monetary aggregates, and it is nowadays admitted that the credibility of the *Bundesbank* stemmed from its inflation-control performance rather than from this seldom-successful strategy. Indeed, money growth was rarely close to the pre-announced target.

When it was established in 1998, the ECB had to emulate the Bundesbank in order to inherit its accumulated credibility. It maintained money targeting while complementing it: Its monetary strategy was initially based on two "pillars," the first pillar being an objective of 4.5% for the annual growth of M3 and the second a combination of leading indicators of inflation such as output prices, import prices, wage costs, etc. Beginning in May 2003, in view of persistent doubts about the stability of money demand in the euro area and of the difficulty for the ECB of meeting its numerical target, the M3 objective was replaced by a broader analysis of monetary trends. But the ECB keeps insisting that money has a "vital role" in monetary policy, while the Fed and other central banks like the Bank of England have stopped giving it a predominant role—or even any role (box 4.22).

Box 4.22 Money and Monetary Policy in the 2000s

In 1943, the Federal Reserve began to regularly publish monetary aggregates. In 1979, it adopted a new strategy that gave a major role to

money targeting. But strict money-supply control was discontinued in the early 1980s. In 2002, the Fed stopped setting targets for those aggregates, and in 2006 it ceased publishing data for M3 altogether, because this aggregate did not appear "to convey any additional information about economic activity that [was] not already embodied in M2 and [had] not played a role in the monetary policy process for many years." By contrast, the ECB maintains that "a model of monetary policy that includes no role for money is incomplete in some important respects" (Trichet, 2006b).

As indicated in box 4.8, standard present-day models of monetary policy do not assign any particular role to money. So what should policymakers do? Surveying the topic, Woodford (2007) found the following arguments for assigning an important role to monetary aggregates in the conduct of monetary policy: A fidelity to the monetarist legacy of the most successful central banks, such as the Bundesbank; the pragmatism involved in cross-checking information, which emphasizes the prominence of the final objective of price stability; and the error-correcting properties of an approach that relies on stocks rather than flows. As developed by Woodford, all three arguments have validity, but there are other ways for a central bank to reach the same goals. Furthermore, giving a prominent role to a concept that is absent from the modern representation of the monetary-policy transmission channels is bound to obfuscate the formulation and communication of the strategy. In fact, because money creation is directly linked to credit growth, monitoring money growth may be thought of as a way to prevent excess lending in the economy, hence to prevent excess risk-taking. This point is developed in section d.

Since the early 1990s, central banks have increasingly relied on a specific implicit rule based on medium-run inflation forecasts—*inflation targeting*. This started in 1990 in New Zealand, soon followed by Canada, the UK, Sweden, and Australia, and similar strategies were thereafter adopted by an increasing number of emerging countries (Brazil, Chile, the Czech Republic, Hungary, Israel, Korea, Mexico, Poland, South Africa, the Philippines, and Thailand). Figure 4.19 gives the inflation bands taken as targets by these various countries. They differ both in levels and amplitudes, which reflects the debate already mentioned on the correct level of inflation.

Inflation targeting is a sophisticated strategy. Contrary to what is commonly believed, it does not target the current rate of inflation, but the central bank's own inflation forecast. This forecast is conditional on all available information, including the current (and possibly future) monetary stance. The transparency of the rule is ensured through publishing both inflation forecasts (frequently accompanied by their standard deviation, as the Bank

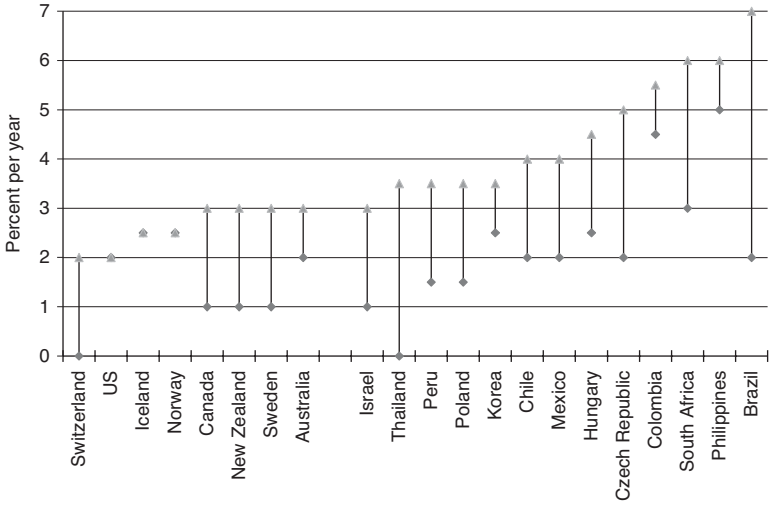


Figure 4.19 Inflation targets in 2005.
Source: Mishkin and Schmidt-Hebbel (2007).

of England routinely does, see figure 4.20) and the models and assumptions used to prepare them. For instance, the Bank of England publishes inflation forecasts based on market interest-rate expectations and on the alternative assumption that interest rates will remain constant. This strategy has several advantages. First, it ensures a high degree of transparency and predictability of monetary policy. Second, it is forward-looking, which allows the central bank to ignore shocks to prices (for example, a rise in the price of oil) as long as they are temporary. Third, it combines the advantages of a rules-based policy with reliance on a wider set of information that is traditionally associated with the discretionary approach. Furthermore, inflation targeting is almost never strictly focused on price stability and puts some weight on stabilizing the real economy.[40]

However, inflation targeting can only work if the central bank is very virtuous and does not use private information (on economic shocks, on transmission channels . . .) for a strategic purpose. This is even more true as central banks gradually abandon the unsatisfactory practice of basing their inflation forecast on a constant interest-rate assumption, to base it on the market forecast of interest rates or even on an explicit pre-announced path for interest rates (in New Zealand, Sweden, and Norway). This requires considerable discipline and sophistication. In some emerging countries, the transparency of inflation targeting may be undermined by the lack of independence of government institutions such as the statistical office.

40. For a recent survey of inflation targeting see Svensson (2008).

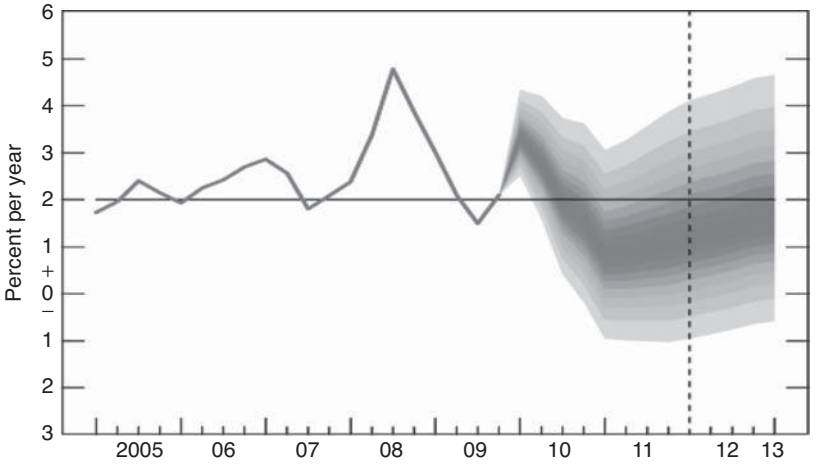


Figure 4.20 Bank of England inflation forecasts, February 2010.
Source: Bank of England.
Note: Each shaded band corresponds to a different confidence interval.

Only time will tell whether inflation targeting represents a significant improvement upon previous monetary strategies. An early assessment (International Monetary Fund, 2006) suggests that its adoption by emerging countries has been associated with both lower inflation and financial market volatility, at no cost in terms of output volatility.

Box 4.23 Central-Bank Transparency

Whatever the monetary strategy, its credibility is enhanced by central-bank transparency. This was defined by Ejffinger and Geraats (2006) as "the extent to which central banks disclose information that is related to the policymaking process." They further distinguish between political transparency (which relates to the information on policy objectives), economic transparency (dissemination of the data, forecasts, and models used by the central bank), procedural transparency (the way monetary decisions are taken), policy transparency (explanation of policy decisions, information on policy inclination, release of minutes, and voting records), and operational transparency (information on errors and unexpected events). Inflation targeting is one way of making monetary policy more transparent in these dimensions. Consistently, the transparency indices constructed by Eijffinger and Geraats credit New Zealand, Sweden, and the UK—three inflation targeters—with high central-bank transparency Table B4.23.1. However, inflation targeting is not a sufficient condition, since Australia (also an inflation targeter) displays lower transparency than the US.

Table B4.23.1
Central bank transparency index, 2002

	Australia	Canada	Euro area	Japan	New Zealand	Sweden	Switzerland	UK	US
Political	3	3	3	1.5	3	3	2.5	3	1
Economic	2	2.5	2.5	1.5	3	2	1.5	3	2.5
Procedural	1	1	1	2	3	3	1	3	2
Policy	1.5	2	2	1.5	3	3	2	1.5	3
Operational	1.5	2	2	1.5	2	3	0.5	2.5	1.5
Total	9	10.5	10.5	8	14	14	7.5	13	10

Source: Eijffinger and Geraats (2006).

c) What reactivity?

A related debate has to do with the frequency and speed of interest-rate adjustments. Two central banks that have adopted the same strategy can react differently to events because they incorporate new information at a different pace and attach a different price to the risk of having to change course. A comparison between the Fed and the ECB illustrates this debate (see figure 4.2). The Fed lowered its interest rates very rapidly to cope with the downturn in 2001–02, and raised them again very quickly in 2004–06 in line with the recovery of the US economy. The ECB reacted much more smoothly to the business cycle during the same period. A somewhat similar pattern was repeated in 2008 in reaction to the financial crisis. The risk of excess activism is to be obliged to change course, to disrupt financial markets, and, due to delays in transmission channels, to end up embarking on procyclical policies. Conversely, the risk with excessive prudence is to fail to exploit the potential of monetary policy as a counter-cyclical policy and thus to increase the burden on fiscal policy, ultimately leading to higher public indebtedness.

The advocates of monetary activism stress that monetary policy can stabilize the economy only if it reacts quickly and vigorously to shocks. By providing private agents with macroeconomic insurance against large business cycles, it enables them to take microeconomic risks. This approach was theorized by Alan Greenspan in 2004:

> The Federal Reserve's experiences over the past two decades make it clear that uncertainty is not just a pervasive feature of the monetary policy landscape; it is the defining characteristic of that landscape. . . . As a consequence, the conduct of monetary policy in the United States has come to involve, at its core, crucial elements of risk management. . . . Policy practitioners operating

> under a risk-management paradigm may, at times, be led to undertake actions intended to provide insurance against especially adverse outcomes.
>
> Alan Greenspan (2004)[41]

The philosophy of the ECB draws on the view that imperfect information about the structure of the economy calls for more inertia and more anti-inflation aggressiveness than in a full-information context (Orphanides and Williams, 2006). For its President Jean-Claude Trichet:

> (The central bank needs) immunizing monetary policy against short-termism by solidly anchoring it on a medium-term perspective. Constantly bombarded by economic news, a central bank risks being swamped by the latest indicator and by its conjectures concerning markets' likely reaction to the latest indicator.
>
> Jean-Claude Trichet (2004)

In other words, the ECB does not want to take the risk of adding its own volatility to market volatility, while the Federal Reserve does not want to take the risk of not having acted in response to a potential threat for economic growth. Furthermore, Trichet (2006a) correctly points out that the mere observation of interest-rate variations does not suffice to determine whether the Fed is more activist than the ECB and suggests that its apparent higher activism may just result from larger shocks hitting the US economy. However, other central banks in Europe, such as the UK and Swedish ones, also change interest rates much more frequently than the ECB and accept having to reverse course within months after a decision. There is more than just events in this difference of behavior. All central banks act on the basis of a precaution principle, as defined in chapter 2, but they disagree on the analysis of which risk would be more harmful to the economy.

d) Is inflation control sufficient for economic stability?

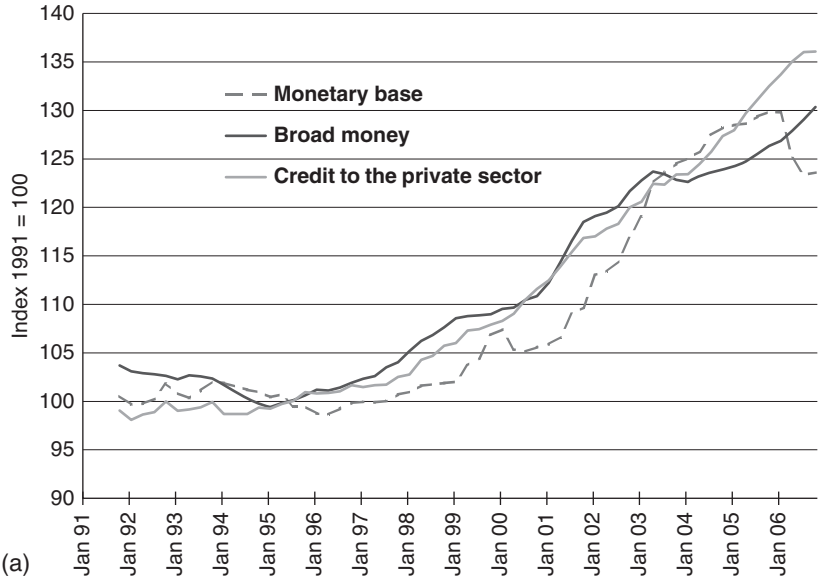
A fourth debate has emerged in the 2000s against the background of ample liquidity creation, asset-price inflation and persistently low price inflation. The issue is whether maintaining price stability can still hold as the main, if not single, objective of the central bank when financial stability is under threat and inflation remains subdued thanks to globalization.

41. In the same speech, available on the Fed Web site, Alan Greenspan takes the example of the Russian crisis of 1998 and of the fall in the Fed official rates in response to this event, which finally appeared for the US economy: "The product of a low-probability event and a potentially severe outcome was judged a more serious threat to economic performance than the higher inflation that might ensue in the more probable scenario."

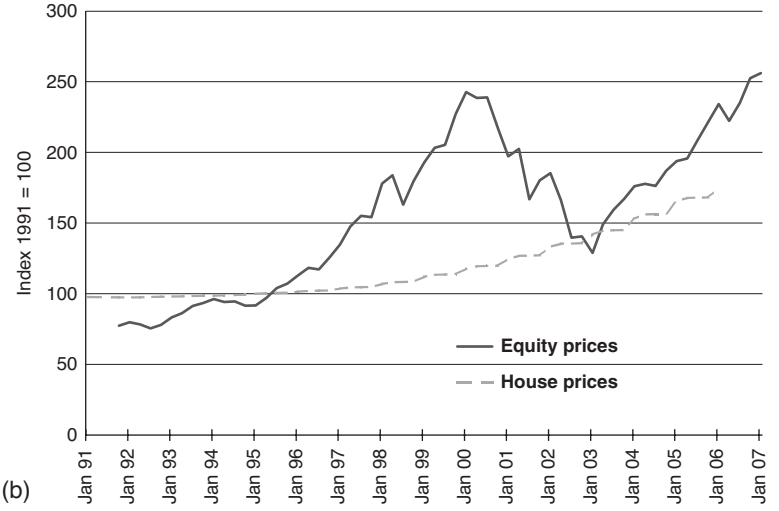
From 1997 to 2007, the ratio of money in circulation over GDP has increased by about one-third in the OECD, equity and house prices have soared, but price inflation has remained close to 2% (figure 4.21). This suggests that instead of giving rise to goods-and-services-price inflation, the increase in liquidity has translated into asset-price inflation.

Whether the central bank should aim at controlling financial-asset-price inflation is not a new issue. After the collapse of the US stock market bubble in 2001, the Fed was criticized for not having raised its interest rates earlier to slow down asset prices in the late 1990s. Similarly, the Bank of Japan did not attempt to avoid the development of an asset-price bubble in the 1980s and the early 1990s but had to cope with the consequences of its bursting. Would it have been preferable to prevent it? It is generally admitted that asset prices (exchange rates, stock and bond prices, property prices) contribute to future developments in the consumer-price index (through the asset-price channel as indicated in section 4.2.2), while the consumer-price index monitored by the central bank measures current prices, not future prices. A forward-looking central bank should therefore take into account indicators of future inflationary pressures.

The issue has gained prominence is the liberalized context of the 2000s. It is increasingly argued that the combination of credible central-bank policies and competitive pressures implied by globalization acts as a powerful break on product-price inflation and thereby reinforces the tendency for inflationary pressures to show up in asset prices instead (Borio, 2006). Though less compelling than it appears, the argument can be formalized in a Barro–Gordon setting (box 4.24).



(a)



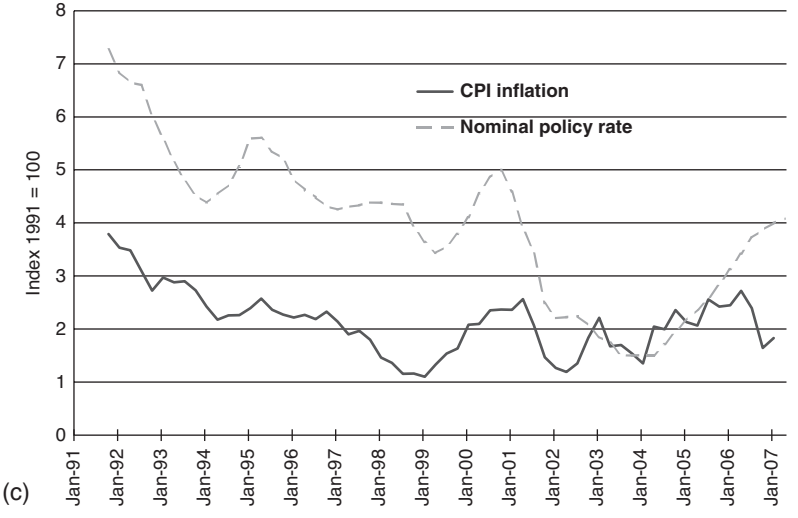


Figure 4.21 Money, asset prices, and price inflation in the OECD, 1991–2007. a) Money and credit, b) asset prices, c) policy rates and inflation.
Source: Bank for International Settlements.
Sixteen OECD countries, weighted averages based on 2000 GDP and PPP exchange rates.

Box 4.24 How Does Globalization Affect Inflation?

In an influential paper published in 2005, Harvard professor Richard Freeman argued that globalization is a huge shock to Western economies since suddenly the global labor pool doubled with China, India, and

the ex-Soviet bloc joining the world economy. This leads to wage moderation worldwide until physical capital accumulates sufficiently for the capital/labor ratio to recover. Any cost-push shock is bound to be short-lived because wages are unlikely to spiral up with prices.

The naïve version of the argument is only superficially convincing, since it in fact confuses relative prices with the general level of prices: Advanced economies could easily sustain higher inflation rates than those of emerging countries if their nominal exchange rate were to depreciate accordingly.

A more rigorous version was offered by Kenneth Rogoff (2003). Drawing on the Barro–Gordon model (box 4.10), he interprets globalization as implying a lowering of the gap k between desired and equilibrium output. The reason is that increased integration reduces the power of insiders and thus lowers the NAIRU. This reduces the incentive for the central bank to inflate; therefore, equilibrium inflation is lower, as indicated by equation (B4.10.2) of box 4.10:

$$\pi_t^{k>0} = \pi_t^{k=0} + \frac{\alpha}{\lambda} k \qquad \text{(B4.24.1)}$$

Furthermore, greater competition in product markets may reduce nominal price rigidities, leading to a higher λ in equation (B4.24.1). This again reduces the equilibrium inflation rate. Both effects, therefore, other things being equal, increase the likelihood that globalization will have the effect of reducing the authorities' incentive to inflate.

Lawrence Ball (2006) has however pointed out that Rogoff's interpretation implies that the Phillips curve should have become *steeper*: A given change in the unemployment rate should correspond to a larger change in the inflation rate (i.e., the inflation cost of expansionary policy has increased). However, the empirical evidence points in the opposite direction: The Phillips curve has generally become *flatter*. This implies that inflation tends to be more stable, but also that incentives to engineer more inflation to decrease the unemployment rate have increased.

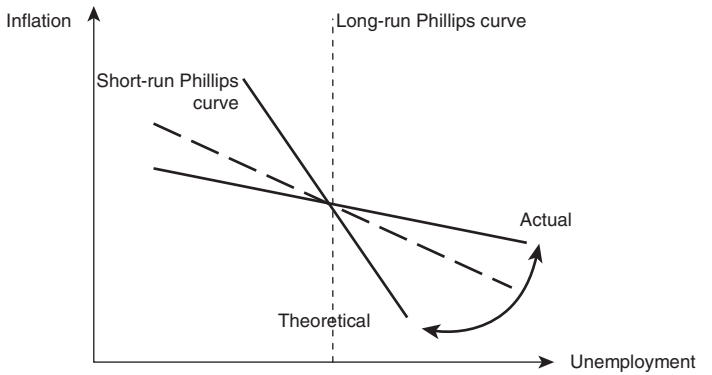


Figure B4.24.1 The effect of globalization on the Phillips curve.

However, central banks generally refuse to target and even to monitor asset prices, due to the difficulty in distinguishing an asset-price bubble from a mere increase in the assets' equilibrium price. In addition, they argue that monitoring asset prices would provide implicit insurance to private investors against the development of asset-price bubbles and create moral hazard, thus excessive risk-taking.[42] According to ECB president Jean-Claude Trichet:

> Experience with past asset price boom episodes tells us that we should be very careful in calling a boom, which is observable, a bubble.... Not all boom or bubble episodes threaten financial stability. Policy-makers should not fall into the trap of attempting to eliminate all risk from the financial system. They would either be unsuccessful (due to moral hazard) or they would likely hamper the appropriate functioning of a market economy where risk-taking is of the essence.
>
> Jean-Claude Trichet (2005b)

A compelling argument against asset-price targeting (Gruen et al., 2003) is that interest-rate movements influence real output with long and variable lags, whereas they burst asset price bubbles immediately due to the forward-looking nature of asset prices. Since the end of a bubble usually causes output to fall due to negative wealth effects, and therefore calls for *lower*, not higher interest rates, the optimal timing for central bank intervention would be to act *ahead* of the asset-price peak. This is hardly possible to guess. As a result, central banks that embark on asset-price targeting run the risk of throwing the economy into a recession when they eventually step in—which is just what the Bank of Japan did in the early 1990s.

Even central bankers who emphasize the specificity of credit-induced asset-price bubbles, such as then Federal Reserve governor Frederick Mishkin (2008), conclude that it is not for monetary policy to prick possible asset-price bubbles, and that the task of mitigating the frequency of such bubbles should primarily be assigned to regulatory policy.

Some central banks, especially the ECB, draw from this discussion the conclusion that money should be rehabilitated as a guide to monetary policy, because inflation targeting leads to focusing excessively on short-term development and neglecting the longer-term risks to economic stability. It is true that money stocks do convey some information on the amount of pent-up inflation that price indicators do not convey. However, this does not make them more reliable.

e) How much ambiguity on liquidity assistance?

We have mentioned that particular financial institutions may be facing funding illiquidity and that certain market segments may be facing

42. See Goodhart and Hofmann (2001) and Bordo and Jeanne (2002).

market illiquidity. In both cases, this may prompt an intervention by the central bank.

It has been recognized at an early stage that *ex post* public intervention is justified when a bank failure would have systemic consequences, either because of the large size of the bank, or because the failure is likely to propagate to others. In such cases, the central banks may provide emergency liquidity to enable the failing bank to refund depositors and other banks that hold creditor positions with this bank. The lender-of-last-resort function was defined in 1873 by British economist and journalist William Bagehot as the ability to lend "without limits, against collateral and at a penalizing interest rate." This possibility both introduces some moral hazard, i.e., banks taking more risks, and attempts at limiting it through setting harsh conditions for liquidity provision.

This raises two issues. The first one is how committed the central bank should be to rescuing financial institutions facing a liquidity shortage. The traditional answer is *constructive ambiguity*,* i.e., intentional ambiguity of the central bank toward the attitude it would have in the case of a financial crisis, in order to have markets behave in a prudent manner. However, this answer is not credible for banks that have reached systemic dimensions, and some central banks such as the Swiss National Bank (which oversees UBS and Crédit Suisse, two banks whose balance sheets dwarf Switzerland's GDP) prefer to spell out in advance what would be the terms of emergency liquidity provision. The Bank of England was also prompted by the Northern Rock episode (see above) to revise its attitude toward moral hazard, as was the Federal Reserve: According to its chairman Ben Bernanke:

> Although central banks should give careful consideration to their criteria for invoking extraordinary liquidity measures, the problem of moral hazard can perhaps be most effectively addressed by prudential supervision and regulation that ensures that financial institutions manage their liquidity risks effectively in advance of the crisis. . . . If moral hazard is effectively mitigated, and if financial institutions and investors draw appropriate lessons from the recent experience about the need for strong liquidity risk management practices, the frequency and severity of future crises should be significantly reduced.
>
> Ben Bernanke (2008)

The second issue is the distribution of the ultimate budgetary cost when the troubled institution eventually requires an injection of fresh capital and when it has a transnational dimension. Unlike temporary liquidity problems, solvency problems require budgetary means and therefore the responsibility for bailing out insolvent banks does not rest with the central banks but with the treasury. The government then *recapitalizes** the bank by purchasing stocks or preferred shares[43] in order to boost its capital and restore the ability of the bank to lend. Public recapitalization was used a number of times in the

43. Preferred shares are senior to common shares but do not carry voting rights.

1990s (in Scandinavia, France, Japan) and in the 2000s (in the US and many European countries).

On the whole, public intervention in banking matters is made very delicate by the interdependence between the legal action by supervisors (bank supervision), fiscal action by governments (recapitalization), and monetary action by central banks (lender of last resort). In theory, bank supervision performs *a priori*, while recapitalization and monetary refinancing act *a posteriori*. In theory again, recapitalization deals with solvency problems (nonperforming loans, capital losses on financial assets held by the bank . . .), while central bank refinancing deals with liquidity problems (when the liquidity of assets is low compared to that of liabilities). However, the distinction between solvency and liquidity is not always easy *ex ante* (exactly as in the discussion on government solvency in chapter 3). It has been proposed that banks be required to prepare and regularly update *living wills** to prepare for possible insolvency and make it more orderly in case it would happen. There is a risk that insufficient banking supervision and the absence of fiscal intervention (recapitalization) will lead the central bank to bear the whole responsibility for the rescue of the banking system, creating a potential conflict with the price-stability objective. There is also a risk that government intervention will shift the burden of supporting banking losses to taxpayers, when it should have been primarily borne by the shareholders of the banks, who have willingly chosen to invest in it and were in full capacity to change its strategy. This raises difficult political economy issues, evidenced in 2008 and 2009 by the discussions between the US administration and Congress on plans to relieve US banks from their impaired assets.

The cross-border operations of international banks add to the complexity, since the operations of overseas branches are supervised both by their *home regulator* (that is, where the headquarters are located) and *host* regulators (that is, where the branch is located), who may or may not agree with each other in case of a liquidity problem. Although committees of supervisors have been established for major international banks, their role is limited to sharing information on risks (which is not necessarily compatible with their role of, and incentives to, protect and promote banks in their jurisdiction. In the 2007–09 crisis supervisors have reportedly not been keen to share information on the vulnerability of banks under their supervision). Governments and central banks are reluctant to engage in *ex ante* discussions on the distribution of responsibilities for assisting transnational financial institutions. In principle, coordination is done in committees. In practice, the issue of international co-ordination of bank supervision has not yet received a satisfactory answer.

References

Akerlof, G., and J. Yellen (1985), "A Near-Rational Model of the Business Cycle with Wage and Price Inertia," *Quarterly Journal of Economics*, no. 100, pp. 823–38.

Akerlof, G., W. Dickens, and G. Perry (1996), "The Macroeconomics of Low Inflation," *Brookings Papers on Economic Activity*, no. 1, pp. 1–76.

Akerlof, G., W. Dickens, and G. Perry (2000), "Near-Rational Wage and Price Setting and the Long Run Phillips Curve," *Brookings Papers on Economic Activity*, no. 1, pp. 1–44.

Alesina, A., and L. Summers (1993), "Central Bank Independence and Macroeconomic Performance: Some Comparative Evidence," *Journal of Money, Credit and Banking*, 25–2, pp. 151–62.

Angeloni, I., and M. Ehrmann (2003), "Monetary Transmission in the Euro zone: Early Evidence," *Economic Policy*, 37, pp. 470–92.

Angeloni, I., L. Aucremanne and M. Ciccarelli (2006), "Price Setting and Inflation Persistence: Did EMU Matter?," *ECB Working Paper Series*, No. 597.

Assenmacher-Wescher, K., and S. Gerlach (2006), "Understanding the Link Between Money Growth and Inflation in the Euro Area," *CEPR Discussion Paper* 5683, May.

Ball, L. (2006), "Has Globalization Changed Inflation?," *NBER Working Paper* 12687, November.

Barro, R. (1995), "Inflation and Economic Growth," *Bank of England Quarterly Bulletin*, May.

Barro, R. (1997), *Determinants of Economic Growth*, MIT Press.

Barro, R., and D. Gordon (1983), "A Positive Theory of Monetary Policy in a Natural Rate Model," *Journal of Political Economy*, 91, pp. 589–610.

Bean, C., J. Larsens, and K. Nikolov (2002), "Financial Frictions and the Monetary Policy Transmission Mechanism: Theory, Evidence and Policy Implications," *ECB working paper*, no. 113, available on the ECB Web site.

Beetsma, R., and H. Uhlig (1999), "An Analysis of the Stability and Growth Pact," *Economic Journal*, 109, pp. 546–71.

Bernanke, B. (2002), "Deflation: Making Sure 'IT' Doesn't Happen Here," Remarks Before the National Economists Club, 21 November, available on the Federal Reserve Board Web site.

Bernanke, B. (2008), "Liquidity Provision by the Federal Reserve," Speech at the Federal Reserve Bank of Atlanta Financial Markets Conference, 13 May.

Bernanke, B. (2009), "The Crisis and the Policy Response," Speech delivered at the Stamp Lecture, London School of Economics.

Bernanke, B., and M. Gertler (1995), "Inside the Black Box: The Credit Channel of Monetary Policy Transmission," *Journal of Economic Perspectives*, 9, pp. 27–48.

Bernanke, B., and M. Gertler (2001), "Should Central Banks Respond to Movements in Asset Prices?," *American Economic Review*, 91, pp. 253–57.

Bernanke, B., and I. Mihov (1997), "What Does the Bundesbank Target?," *European Economic Review*, 41, pp. 1025–54.

Blanchard, O., G. Dell'Ariccia, and P. Mauro (2010), "Rethinking Macroeconomic Policy," IMF Staff Position Note 10/03, February.

Blinder, A. (1997), "What Central Bankers Could Learn from Academics, and Vice-versa," *Journal of Economic Perspectives*, Spring, pp. 3–19.

Blinder, A. (2007), "Monetary Policy by Committee: Why and How?," *European Journal of Political Economy*, 23, 106–23.

Blinder, A., and J. Morgan, (2005), "Are Two Heads Better Than One? Monetary Policy by Committee," *Journal of Money, Credit, and Banking*, 37, pp. 798–811.

Bordo, M., and O. Jeanne (2002), "Monetary Policy and Asset Prices: Does 'Benign Neglect' Make Sense?," *International Finance*, 5, pp. 139–64.

Borio, C. (2006), "Monetary and Prudential Policies at Crossroads," *Bank for International Settlements Working Paper* no. 216, September.
Boskin, M., J. Ellen, R. Dulberger, R. Gordon, Z. Griliches, and D. Jorgenson (1996), *Toward a More Accurate Measure of the Cost of Living*, The Boskin Commission report, final report to the Senate Finance Committee.
Bruno, M., and W. Easterly (1996), "Inflation and Growth: In Search of a Stable Relationship," *Federal Reserve Bank of Saint Louis Review*, 78, pp. 139–46.
Buiter, W. (2006), "Rethinking Inflation Targeting and Central Bank Independence," Inaugural lecture at the London School of Economics, available on W. Buiter's Web site, http://blogs.ft.com/maverecon/.
Caballero, R. (1999), "Aggregate Investment" in Taylor, J., and Woodford, M. (eds.), *Handbook of Macroeconomics*, Elsevier.
Cagan, P. (1956), "The Monetary Dynamics of Hyperinflation," in M. Friedman (ed.), *Studies in the Quantity Theory of Money*, University of Chicago Press.
Calvo, G. (1983), "Staggered Prices in a Utility Maximizing Framework," *Journal of Monetary Economics*, 12, pp. 383–98.
Chinn, M., and G. Meredith (2004), "Monetary Policy and Long-Horizon Uncovered Interest Parity," *IMF Staff Papers*, 51, pp. 409–30.
Clarida, R., J. Gali, and M. Gertler (1999), "The Science of Monetary Policy: A New Keynesian Perspective, *Journal of Economic Literature*, 37, pp. 1661–707.
Crow, C., and E. Meade (2008), "Central Bank Independence and Transparency: Evolution and Effectiveness," *European Journal of Political Economy*, 24, pp. 763–77.
Cukierman, A., S. Webb, and B. Neyapti (1992), "Measuring the Independence of Central Banks and Its Effect on Policy Outcomes," *World Bank Economic Review*, 6, pp. 353–98.
Cunningham, A. (1996), "Measurement Bias in Price Indices: An Application to the UK's RPI," *Bank of England Working Paper*, no. 47, March.
Debelle, G. (2004), "Household Debt and the Macroeconomy," *BIS Quarterly Review*, March, pp. 51–64.
De Larosière, J. (2009), *Report of the High-Level Group on Financial Supervision in the EU*, Brussels, 25 February.
Diamond, D., and P. Dybvig (1983), "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, 91, pp. 401–19.
Dickens, W., L. Goette, E. Groshen, S. Holden, J. Messina, M. Schweitzer, J. Turunen, and M. Ward (2007), "How Wages Change: Micro Evidence from the International Wage Flexibility Project," *Journal of Economic Perspectives*, 21, pp. 195–214.
Dixit, A., and L. Lambertini (2003), "Interaction of Commitment and Discretion in Monetary and Fiscal Policies," *American Economic Review*, 93, pp. 1522–42.
Dornbusch, R. (1976), "Expectations and Exchange Rate Dynamics," *Journal of Political Economy*, 84, pp. 1161–76.
Eijffinger, C., and P. Geraats, (2006), "How Transparent Are Central Banks?," *European Journal of Political Economy*, 22, pp. 1–21.
Fahri, E., and J. Tirole (2008), "Leverage and the Central Banker's Put," paper prepared for the American Economic Association 2009 meeting.
Fama, E. (1984), "Forward and Spot Exchange Rates," *Journal of Monetary Economics*, 14, pp. 319–38.
Fender, I., A. Frankel, and J. Gyntelberg (2008), "Three Market Implications of the Lehman Bankruptcy," *BIS Quarterly Review*, December, pp. 6–7.

Fischer, S. (1977), "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Political Economy*, 85, pp. 191–205.
Frankel, J., and K. Froot (1989), "Forward Discount Bias: Is There an Exchange Risk Premium?," *Quarterly Journal of Economics*, 104, pp. 139–61.
Freeman, R. (2005), "What Really Ails Europe (and America): The Doubling of the Global Workforce," *The Globalist*, 3 June, www.theglobalist.com.
Freixas, X., and J.-Ch. Rochet (1997), *Microeconomics of Banking*, MIT Press.
Friedman, M. (1968), "The Role of Monetary Policy," *American Economic Review*, 58, pp. 1–17.
Friedman, M., and A. Schwartz (1971), *AMonetary History of the United States, 1867–1960*; NBER Studies in Business Cycles no. 12, Princeton University Press.
Goodhart, C., and M. Hofmann (2001), "Asset Prices, Financial Conditions, and the Transmission of Monetary Policy," available on the Reserve Bank of San Francisco Web site.
Gourinchas, P.-O., and A. Tornell (2004), "Exchange Rate Puzzles and Distorted Beliefs," *Journal of International Economics*, 64, pp. 303–33.
Greenspan, A. (1999), Testimony before the Committee on Banking and Financial Services, US House of Representatives, 22 July.
Greenspan, A. (2004), "Risks and Uncertainty in Monetary Policy ," Speech delivered at the American Economic Association annual meeting, available on the Federal Reserve Board Web site.
Gruen, D., M. Plumb, and A. Stone (2003), "How Should Monetary Policy Respond to Asset-Price Bubbles," in A. Richards and T. Robinson (eds.), *Asset Prices and Monetary Policy*, Proceedings of the Research Conference of the Reserve Bank of Australia, pp. 260–80.
HM Treasury (2001), "The Specification of the Inflation Target," and "The U.K. Model of Central Bank Independence: An Assessment," in Balls, E., and G. O'Donnell (eds.), *Reforming Britain' s Economic and Financial Policy*, Palgrave, pp. 71–109.
Hördahl, P. (2009), "Disentangling the Drivers of Recent Shifts in Break-even Inflation Rates," *BIS Quarterly Review*, March, pp. 10–11.
Hume, D. (1742), "Of Interest," and "Of Money," in *Essays, Moral, Political, and Literary*, reedition: Cosimo Classics, 2007.
International Monetary Fund (2006), *Inflation Targeting and the IMF*, available on the IMF Web site.
Ivashina, V., and D. Scharfstein (2010), "Bank Lending During the Financial Crisis of 2008," *Journal of Financial Economics*, forthcoming.
Keynes J.M. (1936), *The General Theory of Employment, Interest and Money*, Macmillan Cambridge University Press.
Kiyotaki, N., and J. Moore (1997), "Credit Cycles," *Journal of Political Economy*, 105, pp. 211–48.
Krugman, P. (1998), "Japan: Still Trapped," available on Paul Krugman's Web site, http://web.mit.edu/krugman/.
Krugman, P. (2000), "Thinking about the Liquidity Trap," *Journal of the Japanese and International Economies*, 14, pp. 221–37.
Lebow, D., and J. Rudd (2001), "Measurement Error in the Consumer Price Index: Where do We Stand?," *Journal of Economic Literature*, 41, pp. 159–201.
Lucas, R. (1972), "Expectations and the Neutrality of Money," *Journal of Political Economy*, 1, pp. 103–24.

Lucas, R. (1996), "Nobel Lecture: Monetary Neutrality," *Journal of Political Economy*, 104, pp. 661–82.

Mankiw, N.G. (1985), "Small Menu Costs and Large Business Cycles: A Macroeconomic Model of Monopoly ," *Quarterly Journal of Economics*,100, pp. 529–39.

McCandless, G., and W. Weber (1995), "Some Monetary Facts," *Federal Reserve Bank of Minneapolis Quarterly Review*, 19, pp. 2–11.

Meier, A. (2009), "Panacea, Curse or Nonevent? Unconventional Policy in the United Kingdom," IMF Working Paper, 09/163.

Mishkin, F. (2008), "How Should We Respond to Asset Price Bubbles?," Speech at the Wharton Financial Institutions Center and Oliver Wyman Institute's Annual Financial Risk Roundtable, Philadelphia, Pennsylvania, 15 May.

Mishkin, F., and K. Schmidt-Hebbel (2007), "Does Inflation Targeting Make a Difference?," NBER Working Paper, 12876.

O'Donnell, G. (2001), "UK Policy Co-ordination: The Importance of Institutional Design," Mimeo, HM Treasury.

Orphanides, A., and J. Williams (2006), "Robust Monetary Policy with Imperfect Knowledge," *Journal of the European Economic Association*, 4, pp. 366–75.

Papademos, L. (2006), "Price Stability, Financial Stability and Efficiency, and Monetary Policy," *BIS Review*, No. 64.

Phelps, E. (1967), "Phillips Curves, Expectations of Inflation and Optimal Employment over Time," *Economica*, 34, pp. 254–81.

Phelps, E. (1973), "Inflation in the Theory of Public Finance," *Swedish Journal of Economics*, 75, pp. 67–82.

Pollard, P. (2004), "Monetary Policy-making around the World: Different Approaches from Different Central Banks," Federal Reserve Bank of Saint Louis.

Rogoff, K. (1985), "The Optimal Degree of Commitment to an Intermediate Monetary Target," *Quarterly Journal of Economics*, 100, pp. 1169–90.

Rogoff, K. (2003), "Globalization and Global Disinflation," *Federal Reserve Bank of Kansas City Economic Review*, Fourth Quarter, pp. 45–78.

Rubin, R., and J. Weisberg (2003), *In an Uncertain World: Tough Choices from Wall Street to Washington*, Random House.

Sargent, T., and N. Wallace (1981), "Some Unpleasant Monetarist Arithmetic," *Federal Reserve Bank of Minneapolis Quarterly Review*, Autumn, pp. 1–17.

Shiratsuka, S. (1999), "Measurement Errors in the Japanese Consumer Price Index," *Bank of Japan, Monetary and Economic Studies*, 17, No. 3, December.

Siebert, A. (2006), "Central Banking by Committees," *DNB Working Paper*, no. 091/2006.

Stiglitz, J., and A.Weiss (1981), "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, 71, pp. 393–410.

Svensson, L. (1999), "Inflation Targeting as a Monetary Policy Rule," *Journal of Monetary Economics*, 43, pp. 607–54.

Svensson, L. (2001), "Inflation Targeting: Should It Be Modeled as an Instrument Rule or a Targeting Rule?," *European Economic Review*, 46, pp. 771–80.

Svensson, L. (2003), "Escaping from a Liquidity Trap and Deflation: The Foolproof Way and Others," *Journal of Economic Perspectives*, 17, pp. 145–66.

Svensson, L. (2008), "Inflation Targeting," *The New Palgrave Dictionary of Economics*, 2nd edition.

Taylor, J. (1980), "Aggregate Dynamics and Staggered Contracts," *Journal of Political Economy*, 88, pp. 1–23.
Tirole, J. (2008), "Liquidity Shortages: Theoretical Underpinnings," in Banque de France, *Financial Stability Review*, Special issue on liquidity, no. 11, February.
Trichet, J.-C. (2004), "Issues in Monetary Policy: Views from the ECB," Speech delivered at the Economic Club of New York, 26 April.
Trichet, J.-C. (2005a), "Monetary Dialogue," Statement before the Committee on Economic and Monetary Affairs, European Parliament, 14 September.
Trichet, J.-C. (2005b), "Asset Price Bubbles and Monetary Policy," Mas lecture, Singapore, 8 June, available on the ECB Web site.
Trichet, J.C. (2006a), "Activism and Alertness in Monetary Policy," Lecture at the Bank of Spain, 8 June.
Trichet, J.-C. (2006b), "Money's Vital Role in Monetary Policy," *Financial Times*, 8 November.
Turner, A. (2009), *The Turner Review: A Regulatory Response to the Global Banking Crisis*, Financial Services Authority, March.
Ueda, K. (2000), "The Transmission Mechanism of Monetary Policy Near Zero Interest Rates: The Japanese Experience 1998–2000," available on the Bank of Japan Web site.
Walsh, C. (1995), "Optimum Contracts for Central Bankers," *American Economic Review*, 85, pp. 150–67.
Woodford, M. (2001), "Fiscal Requirements for Price Stability," *Journal of Money, Credit and Banking*, 33, pp. 669–728.
Woodford, M. (2007), "How Important Is Money in the Conduct of Monetary Policy?," CEPR Discussion Paper no. 6211, March.
Wyplosz, C. (2001), "Do We Know How Low Should Inflation Be?," in Garcia Herrero, A., V. Gaspar, L. Hoodguin, J. Morgan, and B. Winkler (eds.), *Why Price Stability?*, Frankfurt: European Central Bank.

5

International Financial Integration and Foreign-Exchange Policy

5.1 Issues
- 5.1.1 A brief history of the international monetary system
- 5.1.2 Currency convertibility and exchange-rate regimes
- 5.1.3 The foreign-exchange market and the balance of payments
- 5.1.4 Exchange-rate dynamics

5.2 Theories
- 5.2.1 Equilibrium exchange rates
- 5.2.2 Exchange-rate regime choice
- 5.2.3 Models of currency crises

5.3 Policies
- 5.3.1 Capital mobility and the choice of an exchange-rate regime
- 5.3.2 Managing floating exchange rates
- 5.3.3 The future of the international monetary system

References

While monetary policy focuses on the internal value of the currency (the purchasing power of money in terms of goods and services produced locally), exchange-rate policy is concerned with its external value (the purchasing power of money in terms of goods and services produced abroad). The two are intricately related: As seen in chapter 4, the exchange rate is an important channel of transmission of monetary policy and there cannot be any long-lasting divergence between the internal and external purchasing power of a currency. Unless capital movements are tightly controlled, exchange rates cannot be manipulated in an effective way by governments. For many professional economists, the rationale, objective, and choice of instruments of any "exchange-rate policy" are widely debated.

The reluctance of the economic profession to address exchange-rate policy issues contrasts with the passion such issues raise among politicians and in the general public. Every grown-up Briton remembers the infamous

"Black Wednesday" of 1992 when the pound sterling was expelled from the European Monetary System. In Germany, the Kohl administration of the early 1990s came under criticism for having fixed the former East German currency, the ostmark, at par with the D-Mark in the wake of German unification. In many Asian economies, the fixed exchange rates of the 1990s ended in a crisis with devastating economic and social consequences, but there is now talk of moving toward an Asian Monetary Union. China has encountered growing pressures during the 2000s to substantially revalue its currency, which had been fixed in US dollar terms since 1994, except for a 15%, gradual revaluation from 2005 to 2008. When Latvia, a Baltic country, was bailed out by the International Monetary Fund in 2008, it was debated whether its currency, the lat, should remain pegged at a fixed rate against the euro or be allowed to float freely. In 2010, the budgetary crisis in Greece elicited concern that the country might have to leave the euro area. There are many such examples. Choosing the exchange-rate regime or policy are among the major policy decisions a government can take.

5.1 Issues

5.1.1 A brief history of the international monetary system

In medieval Europe, most trade was settled with gold and silver coins. This system evolved gradually, with merchants moving to use paper currency and letters of credit, but confidence in this currency still depended on the possibility of changing it into precious metal. The *Gold Standard**, under which the value of each national currency was determined by a given gold weight, was extended to all major economies in the 1880s and this lasted until World War I. By construction and arbitrage, the Gold Standard involved fixed exchange rates between national currencies.[1] In the second half of the nineteenth century, it provided for an unprecedented expansion of trade. The economic dominance of the British Empire allowed the pound sterling to develop as an international currency parallel to gold, even though the gold reserves of the Bank of England did not cover the whole of the currency issued in pounds sterling.[2] The constraint that this was imposing on macroeconomic stabilization was, however, not fully perceived (see footnote 4). Also, an often understated feature of the pre-World-War-I era was a high degree of capital mobility, especially through the London-based bond market (International Monetary Fund, 1997). Investment needs were concentrated in the railways and government sectors. For sustained periods, Australia, Canada, and Argentina experienced current-account deficits of more than 10% of GDP—financed by foreign (mostly British) capital inflows.

1. In reality, transport and transaction costs introduced a small wedge between bilateral exchange rates and the ratio of the gold values of currencies.
2. The foreign-exchange reserves of the Bank of England were only disclosed in 1931.

World War I put a brutal end to this "first globalization". The war disrupted the international financial market, and destroyed part of the productive capital stock, thus causing inflationary pressures in all Gold Standard countries, reallocated wealth and hence gold stocks across countries, and questioned the role of Great Britain as an economic superpower. Between 1920 and 1924, the exchange rates between the major currencies were largely left to market forces. While the pound had depreciated by 35% against the US dollar between 1913 and 1920, the UK adopted severe deflationary policies that resulted in increasing unemployment. The pound appreciated, and this was further amplified by speculators anticipating a return of the pound to its prewar gold parity, which Churchill, then Chancellor of the Exchequer, eventually decided on in April 1925. Most countries took similar decisions so that by 1927, the Gold Standard had been restored. France joined in 1928, albeit at a depreciated rate, in the wake of the stabilization program led by Raymond Poincaré.

The Gold Standard in the interwar period[3] was not as successful as before and it was finally abandoned in the 1930s. Many reasons explain these mixed results. Given the impact of the war on the UK's economy and balance of payments, the pound was overvalued at its prewar parity. The emergence of New York and of Paris as financial centers competing with London eroded the UK's financial supremacy and the income it received from financial services, which made it more difficult to finance its trade deficit. The available supply of gold was unequally distributed, with the US and France sitting on the biggest stocks while the UK lacked sufficient reserves. This situation was compounded by the behavior of surplus countries which, concerned by inflationary pressures, obstructed the automatic adjustment mechanism.[4] France and the US, in particular, were criticized for not pursuing expansionary policies consistent with their gold assets while the UK was criticized for not restraining domestic credit enough.[5] France also requested the conversion into gold of foreign-exchange reserves accumulated during the Poincaré stabilization. In such a context, the automatic adjustment mechanism that underlaid the proper functioning of the Gold Standard could not work. Finally, the political and economic environment became more fragile, as business cycles were not synchronized and war reparations had an impact on

3. See notably Eichengreen (1992).

4. This automatic adjustment mechanism resulted from one of the so-called "rules of the game" of the Gold Standard (McKinnon, 1993) : Surplus countries, which accumulate reserves, must let their money supply expand, while deficit countries, which lose reserves, must let their money supply contract. Payment imbalances thus naturally translate into price movements that tend to correct the imbalances. See also the discussion on the price-specie flow mechanism later in this chapter. While such an automatic adjustment works in theory, it did not work as well in practice, because countries had many opportunities not to respect the rule, and because the pound was widely used as a reserve currency even though the UK's monetary policy was mainly conducted on the basis of the UK's interests and economic situation. The two other "rules of the game" of the Gold Standard were the fixing of each currency price in terms of gold and the free import and export of gold.

5. Clarke (1967) gives a detailed account of monetary cooperation (or difficulty of) during the interwar period.

balance-of-payments imbalances. Finally, the Great Depression of the 1930s further accelerated the demise of the Gold Standard.

In the wake of the Great Depression, most large countries suspended the convertibility of their currencies into gold. As a consequence, the international monetary system, in the 1930s, was a mix of managed floating exchange rates and fixed rates around anchor countries. The UK was the first country to exit the Gold Standard in 1931,[6] followed by Sweden and other European countries. The US devalued the dollar in 1933,[7] and France and other European countries left the Gold Standard in 1935. In the 1930s, domestic employment became an overwhelming concern, contributing to mounting international tensions as every country tried to shift the burden of internal adjustment to its neighbors through protectionist measures and *competitive devaluations**.

After World War II, priority was given to restoring the *convertibility** of currencies, i.e., the possibility of exchanging them freely to carry out trade transactions but it took more than a decade to reach this goal. The *International Monetary Fund** or IMF was created in 1944 at the Bretton Woods Conference. The IMF, an institution of which almost all sovereign States are members, is tasked with monitoring world payments and helping countries that experience temporary balance-of-payment difficulties to avoid a crisis.[8] The Bretton Woods Conference also established a *Gold Exchange Standard** whereby all currencies were convertible in US dollars at an almost fixed rate, while official US dollar holdings were convertible into gold at a fixed rate of $35 per ounce.[9] The US dollar had thus replaced the pound sterling as the anchor of the international monetary system.

The system was inherently fragile, however. There was a contradiction between, on the one hand, the need to supply a rapidly growing world economy with adequate liquidity and, on the other hand, the need to maintain confidence in the dollar, which implied keeping its issuance in line with the gold reserves of the Federal Reserve.[10] Increasingly, during the 1960s, the United States was issuing more short-term debt in US dollars, initially due to post-war transfers to recovering countries, then to finance the imports needed

6. Outflows of gold had left the UK in a situation in which its stock of available reserves was insufficient to guarantee external convertibility of the pound.

7. The US also imposed controls on gold export, thus suspending de facto the full convertibility of the dollar into gold. The dollar was devalued to $35 per ounce, a rate that was maintained until 15 August 1971 when President Nixon unilaterally suspended convertibility.

8. On the IMF, its role and its early development, see Dam (1982).

9. More precisely, all member countries submitted a par value of their currency expressed in terms of gold or in terms of the US dollar using the gold weight of the dollar effective on 1 July 1944 ($35 per Troy ounce). All exchange transactions between member countries had to take place at a rate that could not diverge by more than 1% from the par value. A member could change the par value of its currency only to correct a "fundamental disequilibrium" in its balance of payment and after consultation with the IMF. Canada left the Bretton Woods monetary system in 1951. Other countries remained part of it until its breakdown, in March 1973.

10. Robert Triffin (1960) and Jacques Rueff (1961) were the first to note this contradiction, later labelled the *Triffin dilemma**.

to sustain the Vietnam war and later as US companies heavily invested abroad. Hence, the volume of official dollar reserves held by monetary authorities expanded and there was an increasing risk that any request, or hint of a request, of conversion of official dollar holdings into gold might trigger a currency crisis due to gold shortage. This risk was enhanced by rising inflation. Toward the end of the 1960s and in the early 1970s, speculation accelerated against the dollar and in favor of gold and strong currencies such as the German mark. Eventually, President Nixon unilaterally suspended the convertibility of the dollar on 15 August 1971. The Smithsonian Agreement signed on 18 December 1971 attempted to safeguard the fixed exchange-rate system by choosing a new fixed parity grid, a wider fluctuation margin and a devaluation of the US dollar. However, the *Bretton Woods system** was finally abandoned in 1973 and the main currencies had to float—a new environment ratified in January 1976 by the Jamaica Agreement, incorporated as the second amendment to the IMF Articles of Agreement.[11]

Facing this new environment characterized by increased exchange-rate instability, the Europeans created in 1972 the "*European Snake*,"* which set the fluctuation margins between European currencies and between these currencies and the US dollar (the reason why it was dubbed the "snake in the tunnel"). The breakdown of the Bretton Woods system in March 1973 removed the "tunnel," whereas the "snake" was transformed into a fully-fledged, stable-but-adjustable exchange-rate system in 1979. This *European Monetary System** (EMS) no longer referred to the dollar. In the EMS, all cross exchange rates had to fluctuate within margins of at most $\pm 2.25\%$ (in some cases $\pm 6\%$) around a central rate. Europe was creating a monetary system of its own on a regional basis, a major breakthrough in post-World-War-II monetary history. The story of the EMS was not a quiet one, however. Between 1987 and 1992, the member countries of the European Community progressively freed up the movements of capital, making it increasingly difficult to maintain fixed exchange rates: Net capital outflows in one country were causing its currency to be sold and thus exerted a downward pressure on its external value. In 1992, the pound sterling and the Italian lira had to exit. In 1993, after a new crisis, the fluctuation margins were widened to $\pm 15\%$. The Maastricht Treaty set the aim of an Economic and Monetary Union "at the latest" on 1 January 1999, allowing the EMS to survive until a single currency, the euro, was eventually created. In 1999, the European monetary union (EMU) was initiated with 11 countries (all the then EU members with

11. Amendments to the IMF Articles of Agreement must be approved by the Board of Governors and become effective when three-fifths of the members, having 85% of the total voting power, have accepted the proposed amendment. There have been three such amendments. The first amendment (adopted in 1968 and ratified in 1969) created the Special Drawing Right in an attempt at increasing world liquidity. The second amendment (the Jamaica agreement) was adopted in 1976 and ratified in 1978. The third amendment, adopted in 1990 and ratified in 1992 enables the IMF to suspend the voting rights of member countries that violate the Articles. A fourth amendment, decided in 1997 to double the liquidity available to the Fund through Special Drawing Rights, was ratified by the US Congress only in 2009.

the exception of Denmark, Greece, Sweden, and the UK). The euro area was subsequently extended to Greece, Slovenia, Cyprus, Malta, and Slovakia.

More than six decades after the Bretton Woods conference, the US dollar remains at the core of the international monetary system. A strong regional monetary cooperation has emerged in Europe and, to a much lesser extent, in Asia and in small groupings of countries (table 5.1), and the role of the euro has developed as a store of value, amounting to 27% of central banks' declared foreign-exchange reserves at the end of 2009 as compared to 18% in 1999. However, the US dollar remains the key currency for international transactions as well as the reference currency for exchange-rate policies in most emerging market economies.

During the period described here, Europe was constantly expressing reservations about floating currencies, whereas the US was relatively content with a floating exchange rate. As for emerging market economies, they have experienced both regimes, with many forms of restrained exchange-rate flexibility. Economic theory makes room for all these choices, depending on each economy's history, structure, and environment.

During the Gold Standard, the fixity of bilateral exchange rates was viewed as a building block of international financial integration. Indeed, from 1870 to World War I, capital flows developed to an extent that has hardly been seen again (see figure 5.1). An open question is to what extent the instability of exchange rates following the breakdown of the Bretton Woods system in 1973 has, in addition to capital controls, been an impediment to financial globalization. Indeed, despite the spectacular surge of capital flows, the world economy is far from being financially integrated. In a famous paper published in 1980, Martin Feldstein and Charles Horioka regressed the investment rate on the savings rate of 16 OECD countries over the period 1960–74. They found that, on average, a one-percentage-point increase in the savings rate was concomitant to a 0.89-percentage-point increase in the investment rate. Hence, over this period, international capital flows added little to domestically financed investments. This counter-intuitive result was later labeled the *Feldstein–Horioka puzzle**. More recent estimations found somewhat lower Feldstein–Horioka coefficients.[12]

The Feldstein–Horioka puzzle has been much discussed in the economic literature, mainly in three respects. First, it has been argued that standard econometric techniques are not appropriate to deal with nonstationary saving-and-investment rates. Subsequently, *cointegration** methods have been used

12. For instance, Obstfeld and Rogoff (1995) found a coefficient of 0.69 for a sample of 22 OECD countries over the 1982–91 period and Blanchard and Giavazzi (2002) obtained a 0.58 coefficient for the 30 OECD countries over the 1975–2001 period. Interestingly, they show the coefficient to be lower and declining in the euro area: Only 0.14 over the 1991–2001 period, down from 0.41 over 1975–90.

Table 5.1
Exchange-rate regimes, IMF classification, April 2008

Exchange-rate regime	Number of countries	Countries
No separate legal tender: 40	7 US dollarization	Ecuador, El Salvador, Marshall Islands, Micronesia, Palau, Panama, Timor-Leste
	2 euroization	Montenegro, San Marino
	1 Australian dollarization	Kiribati
Monetary union: 35	15 euro area	Austria, Belgium, Cyprus, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Malta, The Netherlands, Portugal, Slovenia, Spain. Floating exchange rate against the rest of the world
	8 WAEMU[a]	Benin, Burkina-Faso, Ivory Coast, Guinea Bissau, Mali, Niger, Senegal, Togo. Fixed exchange rate against the euro
	6 CAEMC[b]	Cameroon, Central African Republic, Chad, Republic of Congo, Equatorial Guinea, Gabon. Fixed exchange rate against the euro
	6 ECCU[c]	Antigua and Barbuda, Dominica, Grenada, St Kitts and Nevis, St Lucia, St Vincent and the Grenadines. Fixed exchange rate (currency-board type) against the US dollar
Currency board: 7	4 against the euro	Bosnia and Herzegovina, Bulgaria, Estonia, Lithuania
	2 against the US dollar	Djibouti, Hong Kong SAR
	1 against the Singapore dollar	Brunei Darussalam
Conventional fixed pegs: 54	6 against the euro	Cape Verde, Comoros, Croatia, Denmark, Latvia, FYR of Macedonia
	36 against the US dollar	Angola, Argentina, Aruba, Bahamas, Bahrain, Bangladesh, Barbados, Belarus, Belize, Eritrea, Guyana, Honduras, Jordan, Kazakhstan, Lebanon, Malawi, Maldives, Mongolia, Netherlands Antilles, Oman, Qatar, Rwanda, Saudi Arabia, Seychelles, Solomon Islands, Sri Lanka, Surinam, Syria, Tajikistan, Trinidad and Tobago, Turkmenistan, United Arab Emirates, Venezuela, Vietnam, Yemen, Zimbabwe

(*Cont'd*)

Table 5.1
continued

Exchange-rate regime	Number of countries	Countries
	3 against the South African rand	Lesotho, Namibia, Swaziland
	2 against the Indian rupee	Bhutan, Nepal
	7 against a basket	Libya, Fiji, Kuwait, Morocco, Russian Federation, Samoa, Tunisia
Pegged rates with a horizontal band: 3	1 against the euro	Slovak Republic
	2 against the US dollar	Syria, Tonga
Crawling pegs: 10	7 against the US dollar	Bolivia, China, Costa Rica, Ethiopia, Iraq, Nicaragua, Uzbekistan
	3 against a basket	Azerbaijan, Botswana, Iran
Managed floating: 44	8 Against the US dollar	Cambodia, Kyrgyz Republic, Lao PDR, Liberia, Mauritania, Mauritius, Myanmar, Ukraine
	3 against a basket	Algeria, Singapore, Vanuatu
	33 undefined	
Independently floating: 26		Albania, Australia, Brazil, Canada, Chile, Democratic Republic of Congo, Czech Republic, euro area, Hungary, Iceland, Israel, Japan, Repuplic of Korea, Mexico, New Zealand, Norway, Philippines, Poland, South Africa, Somalia, Sweden, Switzerland, Turkey, UK, US, Zambia

[a] West-African Economic and Monetary Union.
[b] Central African Economic and Monetary Community.
[c] East Caribbean Currency Union.

Source: Adapted from International Monetary Fund Web site, www.imf.org/external/np/mfd/er/2008/eng/0408.htm

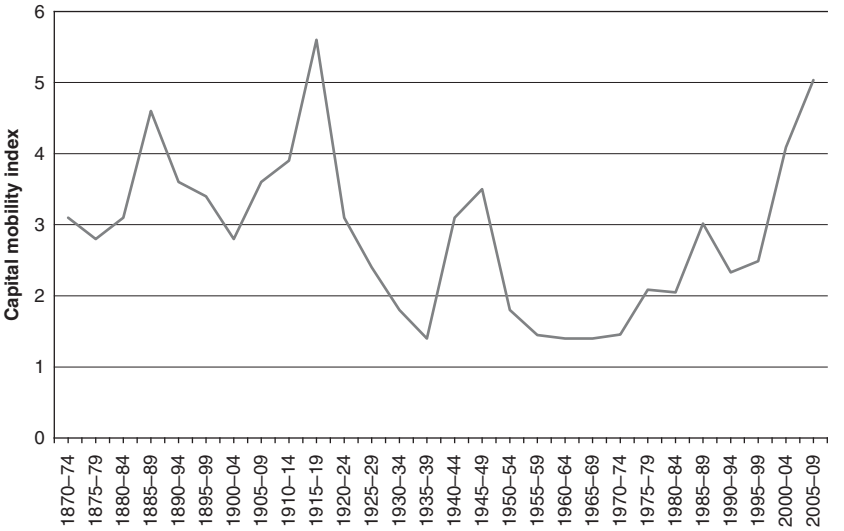


Figure 5.1 International financial integration, 1870–2007.
Source: Taylor (1996), updated by the authors as from 1970 based on OECD data.

Note: The capital mobility index is defined as the average of the absolute values of currents accounts relative to GDP for major capital-importing and capital exporting countries. Countries include Argentina, Australia, Canada, Denmark, France, Germany, Italy, Japan, Sweden, the UK, and the US.

in the literature, leading to a more mixed picture of the saving–investment relationship.[13] Second, it has been noted that the close saving–investment relationship is the result of the intertemporal budget constraint (the saving–investment gap cannot grow forever) or of the endogenous reaction of public authorities (see, e.g., Summers, 1988) which, for instance, raise public saving when investment is buoyant. Finally, the Feldstein–Horioka puzzle can be related to the empirical evidence of a strong *home bias** in portfolio choices: Savers seem to hold fewer foreign assets than optimal portfolio diversification would suggest.[14] The reason for the home bias is still an open question, but information asymmetries are good candidates: Savers have more information on domestic risks than on foreign risk. This reduces their willingness to hold foreign assets. One major counter-example has been the willingness

13. Cointegration is a technique used in time-series analysis. An economic time-series $\{x_1, x_2, \ldots\}$, where the subscript denotes the date of observation, is said to be nonstationary if it does not tend to return to some constant value or deterministic trend after a shock. A random walk, i.e., the cumulated sum of uncorrelated, identically distributed random variables, is an example of a nonstationary time series. Two nonstationary series are said to be "cointegrated" if there exists a linear combination of them that is stationary. When two nonstationary series are correlated but not cointegrated, the correlation is said *spurious*.*

14. Optimal portfolio choice is discussed in section 5.2.

of investors all over the world to hold US assets, thus financing huge US current-account deficits in the 1990s and 2000s.

5.1.2 Currency convertibility and exchange-rate regimes

Not all countries participate in the global financial system. Some countries retain inconvertibility or limited convertibility of their currencies. When the domestic currency is at least partially convertible, its price can either be set freely by the market or managed by the government and the central bank. Governments need to make two crucial decisions: They must decide on the conditions for exchanging the domestic currency for foreign currencies—*currency convertibility**—and they must decide on the extent of exchange-rate flexibility—the *exchange-rate regime**. In the words of chapter 1, the exchange-rate regime is the institutional setup in which the exchange-rate policy is operated. It is both a legal and an empirical notion. Countries have to declare to the IMF both the extent of convertibility and the nature of the exchange-rate regime.

a) Currency convertibility

In most countries, it used to be the case that government would decide the value of the exchange rate[15] against foreign currencies and would submit any foreign-exchange transaction to prior authorization. The currency was then said to be not convertible. This was the case for Western European countries before 1958, for the former Soviet bloc before 1990, and for numerous developing countries until recently. In all other cases, the currency is said to be convertible. It may, however, be convertible for some transactions and nonconvertible for others. More precisely, it is useful to distinguish:

- *Current-account convertibility**: The national currency can be exchanged freely for the purpose of importing goods and services, as well as for current transfers and factor income. This is the case in a majority of countries.
- *Financial-account convertibility**[16]: Direct investments, portfolio investments, and bank loans are permitted without restriction. It is synonymous with *capital mobility* and may concern some financial transactions, partially or totally, and not others. Capital is actually never fully mobile, as there are always valid reasons for control (for example, the fight against money laundering and terrorist finance). However, most advanced economies have liberalized capital flows in the 1980s and

15. There could be more than one exchange rate: for exports and for imports, or depending on product types, or on individuals. In such cases, one would speak of a *dual** or a *multiple exchange-rate system.**

16. The financial account, a balance-of-payment item (see below), was formerly known as the "capital account."

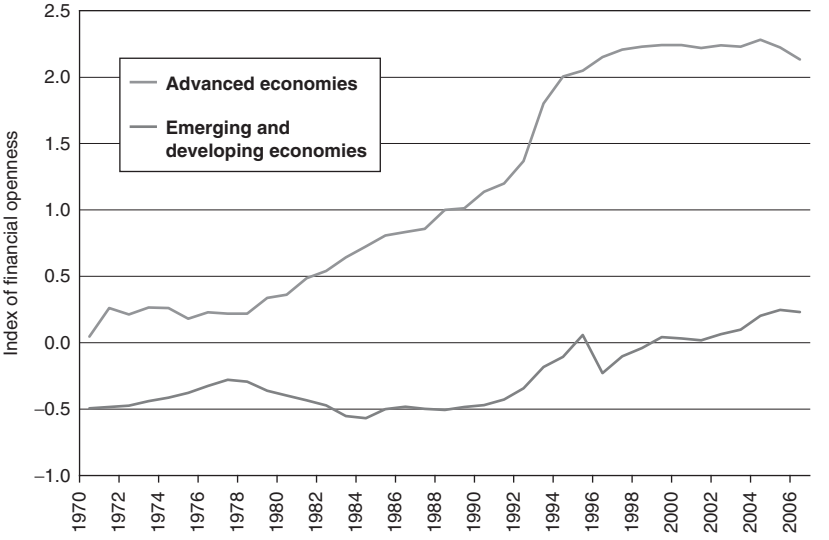


Figure 5.2 Financial openness over time, 1970–2007.
Source: Chinn and Ito (2008), based on the IMF *Annual Report on Exchange Arrangements and Exchange Restrictions.*

Note: The index is computed as the unweighted average of 24 advanced economics and 128 emerging and developing economies.

early 1990s. The movement is more recent and still incomplete in emerging and developing countries (see figure 5.2).

The extent of currency convertibility has important consequences for the determination of exchange rates. When capital does not move freely across countries, foreign-exchange transactions arise only as the counterpart of an underlying "real world" transaction such as exports, imports, or income repatriation. When capital movements are free, foreign-exchange transactions also arise from financial asset purchases and sales, which prove to be much larger and volatile than "real world" transactions, possibly leading to more exchange-rate instability.

b) Exchange-rate regimes and currency crises

Large exchange-rate fluctuations are a major source of uncertainty for the "real world" because of their impact on relative prices across countries, and hence on competitiveness and returns. They also affect the relative value of assets and liabilities. In developing countries, foreign liabilities are generally denominated in key foreign currencies. Hence, a depreciation of the domestic currency raises the value of the external debt. For all these reasons, governments may wish to reduce the extent of exchange-rate fluctuations

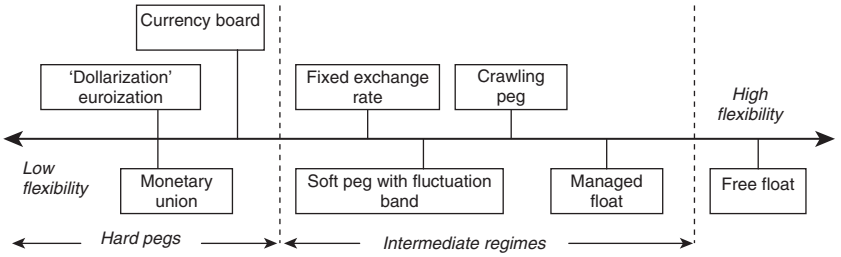


Figure 5.3 Taxonomy of exchange-rate regimes.

or to curb exchange-rate variations. Figure 5.3 ranks exchange-rate regimes depending on the degree of government intervention. The highest degree of fixity is obtained when the national currency is abandoned. A first option is to use the currency of another, larger country. Panama and Ecuador have thus "*dollarized**" their economies, while Montenegro and San Marino have "*euroized**" theirs. A dollarized country becomes unable to control the domestic money supply and its central bank is left with the sole mission of performing technical tasks. Another option is for a group of countries to merge their currencies and create a new one within a *monetary union**. This is the case in the euro area and in African monetary unions. The difference between dollarization and a monetary union stems from the monetary-decision process: There is no Ecuadorian representative sitting at the Federal Open Market Committee that sets the US interest rates, nor is there an official from Montenegro at the Governing Council of the European Central Bank.

A weak form of dollarization is the *currency board**. The national currency continues to circulate but it can be exchanged at a fixed rate against the currency of some larger country, say, the US dollar, or against a basket of currencies. For this purpose, the issuance of domestic currency by the central bank is backed by foreign assets only. Hence, the domestic monetary base is matched by the corresponding amount of foreign currency held by the central bank. This tight rule allows the fixed exchange rate to be credible. It has been used by countries which had lost their monetary credibility following a period of hyper-inflation, such as Argentina in the 1990s. The breakdown of the Argentine currency board in January 2002 shows, however, that a currency board does not offer the same guarantee on the fixed exchange rate as the one provided by dollarization or a monetary union.[17]

*Hard pegs** are exchange-rate regimes in which the fixity of the exchange rate is backed by a tight institutional scheme. This clearly includes currency boards, and, by extension since they do not formally involve any "peg," regimes such as dollarization and monetary union in which there is no separate national legal tender.

17. There are, however, also examples of monetary unions coming to an end, e.g.. the nineteenth-century Latin Union in Europe and the former Soviet Union.

When the fixed exchange rate is not enshrined in legally binding arrangements such as a currency board, its credibility rests primarily on the government will. This is what European countries experienced in the 1980s and 1990s. In the European Monetary System, even though the system had been conceived as a fixed but adjustable peg, a *devaluation** (i.e., downward adjustment of the reference rate around which the market exchange rate was allowed to fluctuate) bore a political cost. The finance minister would return from Brussels, where such decisions were discussed, having "impoverished" his fellow citizens and having defaulted on a European commitment. Such political incentives, however, are not always effective. In the early 1980s, the European monetary system experienced numerous devaluations. In 1992 and 1993, after capital flows were fully liberalized in the European single market, the system had to face two major crises. The vulnerability of conventionally fixed exchange-rate regimes derives from the strength of speculative attacks when the firmness of the commitment is being questioned by markets and when the scope for official intervention is limited (see section 5.2).

Fixed exchange-rate regimes can be given more flexibility by allowing wider fluctuation margins or, in the case of *crawling pegs**, by partially adjusting the rate of devaluation to the inflation differential with trade partners, so that the resulting real overvaluation remains limited while promoting domestic disinflation. More generally, *soft pegs** or *intermediate exchange-rate regimes** cover pegged exchange rates in a wide sense: Fixed or crawling, with or without a fluctuation band or even managed floats (see below). Pegs can be against a single currency (typically, the US dollar or the euro) or against a basket of currencies such as the IMF unit of account, the *Special Drawing Right* (*SDR*)*[18]

Lastly, exchange rates can float more or less freely, without a reference target. The government may intervene occasionally to control the level or variability of the exchange rate, but without setting any nominal objective. Depending on the frequency of such interventions, one will speak of *managed floating** or *free floating**.

Table 5.1 breaks down IMF member countries according to their exchange-rate regimes in April 2008. Of the 188 countries or zones under review, 52 were running a hard-peg regime, i.e., dollarization/euroization, monetary union, or a currency board. Among them, 36 used the euro as their reference currency (including euro area member states) and 14 used the US dollar. At the other end of the spectrum, only 26 currencies (including the euro) were floating freely. In between, 111 countries were running intermediate regimes, i.e., conventional fixed pegs fixed pegs with horizontal bands, crawling pegs, or managed floats. Therefore, the present international monetary system has often been qualified as a "dirty floating" one, where only key currencies are allowed to float whereas the flexibility of smaller currencies is often limited through a wide range of arrangements.

18. The number of countries that peg their currencies at a fixed rate to a basket shrank from 36 in 1990 to only 7 in 2008, see table 5.1.

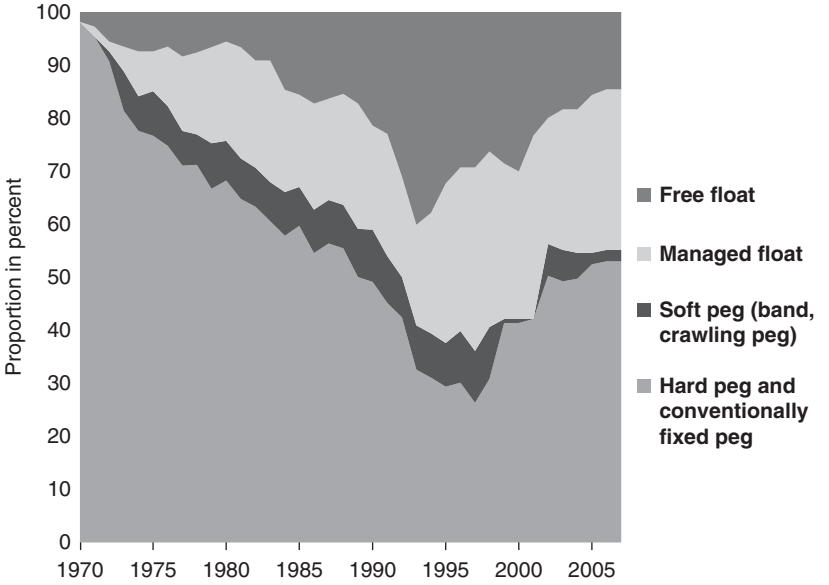


Figure 5.4 Fear of floating: Distribution of exchange-rate regimes, 1970–2007. Source: Ilzetzki et al. (2008) based on IMF data.

Intermediate exchange-rate regimes fell into disgrace after the currency crises of the late 1990s. It was then felt that such regimes were unsustainable in a world of perfect capital mobility. In such a world, the only way to keep a stable exchange rate is to surrender monetary independence and adopt a hard peg. Monetary independence is workable within a free floating regime, but table 5.1, above, shows that this regime was only chosen by 13% of countries in 2008.

Figure 5.4 further evidences the attractiveness of fixed exchange rates: After a long decline from the early 1970s to the late 1990s, fixed exchange-rate regimes (either conventional or hard) did gain in popularity at the expense of free floats, especially in emerging and developing economies where the "fear of floating" is widespread (Calvo and Reinhardt, 2002).

However, conventional pegs are vulnerable: In a world with free capital flows, the foreign-exchange-market participants who expect a currency to be devalued can borrow in this currency on the money market and sell the domestic currency against foreign currencies. Following the law of supply and demand, the price of the domestic currency (which is sold) falls against foreign currencies (which are purchased): The currency is bound to be devalued. The likelihood of such a *speculative attack** depends on the credibility of the government, which can intervene on the foreign-exchange market to buy the domestic currency, or raise the domestic interest rate to make speculative attacks more costly. However, official intervention is limited by the stock of official reserves, and raising the interest rate to face a speculative

attack involves a high cost. Therefore, a speculative attack often leads to a *currency crisis**, as in 1992–93 in Europe, 1994 in Mexico, 1997 in Asia, and at the turn of the millennium in Brazil, Russia, Turkey, and Argentina. These crises involve large macroeconomic costs. For instance, the Asian crisis of 1997 led to a cumulated production loss estimated at 25, 33, and 62% in Korea, Malaysia, and Indonesia, respectively, after three years.[19] The risk of currency crises needs to be balanced against the benefit of a fixed exchange rate in terms of lower inflation or stable debt-service burden (see section 5.2).

5.1.3 The foreign-exchange market and the balance of payments

In countries where the financial system is not developed, or where convertibility is restrained, the domestic currency is exchanged against foreign currencies either by official authorities or in informal markets, and there can be as many exchange rates as there are bilateral transactions. In most countries, however, there is a single market where currencies are exchanged, the *foreign-exchange market**.

a) The foreign-exchange market

The foreign-exchange market is a wholesale market where only financial intermediaries, large corporations, and central banks intervene. Operations take the form of transfers between cash accounts expressed in different currencies. The exchanged good is indeed money (the "M1" monetary aggregate of chapter 4).

If the market is sufficiently active, *arbitrage** between currencies ensures at every point in time the uniqueness of the exchange rate and the transitivity between exchange rates. For instance, if a euro is worth 1.2 US dollars and a US dollar worth 110 Japanese yen, then a euro should be worth $1.2 \times 110 = 132$ yen—otherwise, one could make an easy profit by changing money in all three currencies in a row. Arbitrage is never perfect, due to the difference between the sale price and the purchase price on the market, the *bid–ask spread**, which represents the fee paid to financial intermediaries and depends on the liquidity of the market, i.e., on the frequency and size of transactions. Apart from this fee, there is a single price for a specific currency all around the world at any time, and the matrix of all bilateral prices is consistent. However, the price of a currency depends on the delivery date: For *spot** transactions, the currency is delivered within 24 hours, whereas *forward** transactions imply a delivery at a deferred date.

Based on the spot and on the forward markets, a large number of foreign-exchange *derivatives** have developed. For example, a *foreign-exchange swap** is an exchange of cash flows denominated in two different currencies over a

19. See Mussa et al. (2000).

certain period of time. A *call option** gives its owner the right, but not the obligation to purchase foreign currency at a rate agreed upon beforehand and at a certain date in the future (or, depending on the nature of the option, at any time before this date), while a *put option** gives the right to sell it. Foreign-exchange derivatives can be agreed on over-the-counter, but some of the most standard formats are traded in the marketplace. It can be shown that the value of an option increases with the volatility of the exchange rate over its lifetime (see Garman and Kohlhagen, 1983): The market valuation of options thus provides a way to gauge expected foreign-exchange volatility, the *implicit volatility**.

The foreign-exchange market went through a considerable expansion in the 1990s as a result of three forces: The development of international trade, capital-movement liberalization, and new financial techniques to manage financial risk. The *Bank for International Settlements** (BIS), a public institution based in Basel (Switzerland), is tasked with providing services to central banks and monitoring global financial markets. According to the BIS, foreign-exchange transactions amounted to 3200 billion dollars per day in April 2007, of which 1005 billion were spot transactions. This roughly represented 18 days of world production and 46 days of international trade.

Since 1945, the central role of the US dollar as a *vehicle currency**, i.e., as an intermediary for transactions between third currencies, has not been challenged. In April 2007, only 12% of foreign-exchange transactions did not involve the dollar, while 63% of transactions did not include the euro. It is easier and less expensive to exchange Korean won against dollars, then dollars against Mexican pesos, than to directly exchange won against pesos. This does not imply that the foreign exchange market is based in the US, since transactions are not physical. Indeed, the main center for foreign exchange transactions is London (34%), followed by New York (17%), and Tokyo (6%). Neither does it mean that the dollar has remained unchallenged for other monetary functions. Indeed, since its introduction in 1999, the euro has rapidly emerged as a second international store of value, although not quite yet as a means of payment nor as a unit of account.

b) Balance of payment equilibrium

When the financial account is not convertible, the exchange rate can more easily be fixed by way of administration. However, when companies and households are free to buy and sell foreign assets, the exchange rate has to be constantly consistent with market equilibrium. When it is flexible, it adjusts so as to achieve this equilibrium. When it is fixed, the market cannot be cleared unless the central bank intervenes by selling or buying foreign currency.

In all cases, the relevant instrument to identify supply and demand of foreign currency is the *balance of payments**, which describes all transactions with the rest of the world. The balance of payments is made up of three sections, or "accounts".

The current transaction account or *current account** is the operating account of the country. It describes all earnings from the rest of the world deriving from exports of goods and services, labor and capital income like dividends or remittances, and other current transfers. Symmetrically, it reports all payments made to the rest of the world related to imports of goods and services, labor and capital income, and other current transfers.

The *capital account** tracks capital transfers without a counterpart, like debt forgiveness and investment grants.

The *financial account**, formerly known as the "capital account," describes all sales of domestic assets—private and public securities, borrowings, real estate—to the rest of the world (*capital inflows**) and all purchases of foreign assets (*capital outflows**). When an equity investment results in a share higher than 10% in a foreign company, which allows the exercise of effective control, the investment is called a *foreign direct investment**. Another example of capital inflows and outflows are sales and purchases of foreign securities by the domestic central bank as part of its *foreign-exchange reserves** management. By definition, foreign-exchange reserves are made of foreign-denominated securities and deposits (plus gold) held by the central bank. International sales and purchases of securities that are not classified as foreign direct investments or foreign-exchange reserves are called *portfolio investments**. The financial account also includes a category called *other investments** that primarily includes bank credit.

Since all transactions have to be financed, the net surplus of all three accounts should add up to zero. But all items are not well measured and there are usually substantial statistical discrepancies.[20]

Table 5.2 summarizes the balance of payments of the US and the euro area in 2008. The US economy registered a large current-account deficit as it imported more goods and services than it exported. This deficit was financed by foreign purchases of US stocks and bonds (including foreign, notably Asian, central banks accumulating reserves). In the euro area, in contrast, trade was broadly balanced and inflows and outflows of capital roughly cancelled each other: Foreigners bought European securities and foreign banks lent to European banks and corporations, but euro area investors also invested in noneuro (mainly US and UK) companies.

Figure 5.5 sketches the balance of payments of a country which imports more goods and services than it exports. For the sake of simplicity, we assume a balanced capital account and we omit it in our sketch. This country spends more abroad than it earns from abroad on goods and services, and therefore has to borrow from nonresidents or to sell them financial assets. In a floating-rate regime, the exchange rate can adjust to balance the current

20. To judge the quality of the statistical system, statistical discrepancies should be compared to *gross*, not to *net* figures. In the euro area, for instance, they were larger than the current-account surplus in 2008 (in absolute value) but only represented 5.7% of the gross credit side of the current account.

Table 5.2
The US and euro area balance of payments in 2008

	US		Euro area	
	$bn	% GDP	€bn	% GDP
Current account	**−673.3**	**−4.7%**	**−67.3**	**0.7%**
Goods and services	−681.1		47.0	
Factor income	127.6		−22.0	
Transfers	−119.7		−92.3	
Capital account	**−2.6**	**−0.0%**	**13.7**	**0.1%**
Financial account*	**546.6**	**3.8%**	**212.6**	**2.3%**
Direct investments	7.4		409.2	
Portfolio investments	154.4		235.7	
Financial derivatives	−373.9		−12.3	
Other investments	342.2		102.1	
Foreign exchange reserves	416.5		−4.9	
Statistical discrepancies	**129.3**	**0.9%**	**−151.1**	**−1.6%**

Note: Financial account: net capital inflows (+), net capital outflows (−).

Source: European Central Bank and US Bureau of Economic Analysis.

account and the financial account: If nonresident demand for domestic assets is initially insufficient, for instance, then the exchange rate depreciates to make these assets more attractive. In a fixed-exchange-rate regime, the exchange rate is not allowed to depreciate and the central bank has to step in to clear the market. Practically, it has to sell assets out of its foreign-exchange reserves.

Central bank interventions and domestic monetary policy are closely interrelated. When the central bank sells foreign-exchange reserves, it reduces domestic money supply, since the cash it collects in exchange is withdrawn from circulation. Conversely, an increase in foreign reserves expands money supply. This in turn affects the domestic interest rate. This channel also explains why the impact of monetary policy decisions in a fixed-rate regime is contradicted by reserve accumulation or decrease (this is the case in the Mundell–Fleming model presented in chapter 3). In order to offset the impact of foreign-reserve movements on money supply, the central bank can choose to *sterilize** foreign-exchange interventions by buying or selling Treasury bills in the open market, or to issue short-term paper (*sterilization bonds**) for this particular purpose. As an example, the People's Bank of China sterilizes foreign-exchange reserve accumulation by issuing so-called "PBoC bills" and selling them to Chinese banks. By so doing, the central bank incurs a net budgetary cost if the yield on domestic bills is higher than the yield on

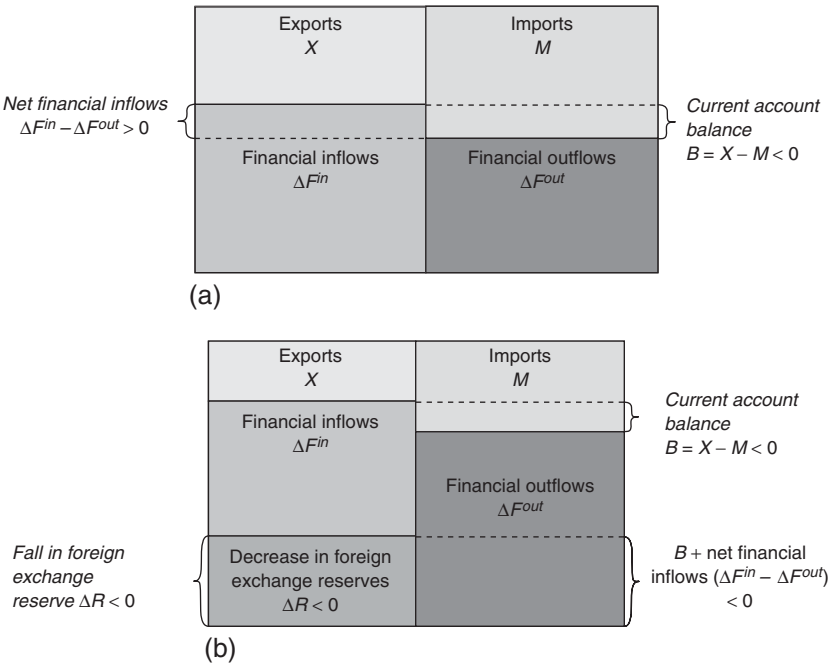


Figure 5.5 The balance of payments of a capital importing country. a) Floating exchange-rate regime; b) fixed exchange-rate regime.
Note: For the sake of simplicity, the capital account is supposed to be balanced and is not represented here. We also neglect factor income as a component of the current account.

foreign-currency denominated reserves. This cost is usually passed to the government. Intervention sterilization will be discussed further in the last section of this chapter.

When a country accumulates current-account deficits, it spends more than it earns one year after another, and therefore builds up liabilities toward the rest of the world. Practically, this means that foreigners hold an increasing amount of the country's domestic securities, private and public; they increase their stakes in domestic companies; or they extend more loans to domestic agents. When a country has more external debt than external assets, it has a negative *net foreign asset position**. This is typically the case of the US. Figure 5.6 shows how the net foreign position has deteriorated since the mid-1980s, while Japan, in contrast, was accumulating assets abroad. The dynamics of net foreign positions also depend on movements in asset prices. In Figure 5.6, the temporary fall in Japan's net foreign assets in 1998 can be ascribed to a rise in the value of the yen relative to the other currencies. Such valuation effects play an important role in current-account adjustment; we shall come back to them later on.

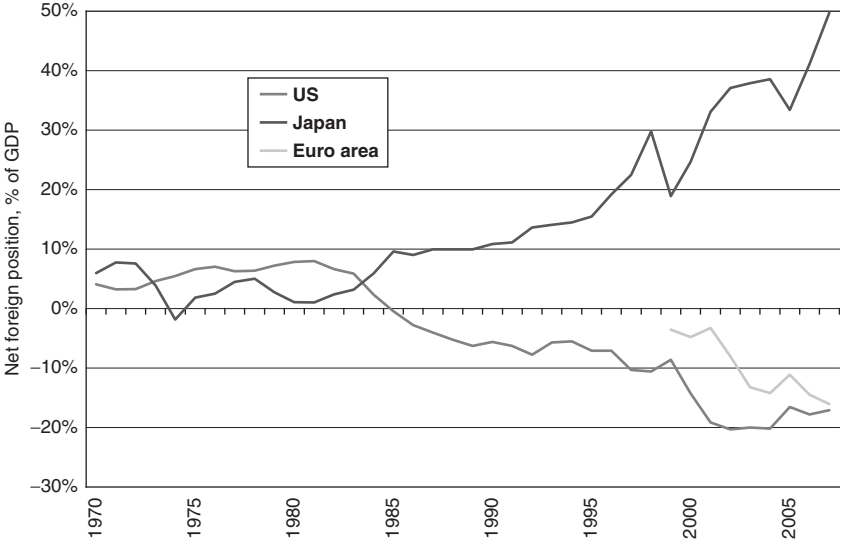


Figure 5.6 Net foreign positions of the US, euro area and Japan, 1970–2007.
Source: External Wealth of Nations database developed by Lane and Milesi-Feretti (2007).

Note: Assets and liabilities are measured at estimated market value.

5.1.4 Exchange-rate dynamics

Figure 5.7 shows the monthly value of the euro versus the US dollar between 1980 and 2010. The figure shows very large fluctuations upward and downward. It is not rare to see the euro appreciate or depreciate by over 10% in just a few months. Such patterns are not specific to this particular couple of currencies, nor to the period under review. High instability has been a characteristic of floating exchange rates since the demise of the fixed-rates system in 1973. Furthermore, exchange-rate volatility has been higher by an order of magnitude than the volatility of macroeconomic variables, which did not increase after 1973 (Flood and Rose, 1995).

Up to this point, we have only discussed the relative value of two currencies, the *nominal exchange rate**. But economists are not concerned chiefly with nominal exchange rates. What matters for consumers and companies' decisions is the relative price of goods, services, and assets, and any movement of the nominal exchange rate can be offset by a price variation, leaving relative prices unchanged. To assess price competitiveness, one has therefore to correct the observed exchange rate with relative prices: This is called the *real exchange rate**. Moreover, in the economy as a whole, appreciation with respect to a trading partner can be offset by depreciation with respect to another. A synthetic image of the competitiveness of a country with respect to the rest of the world can be obtained by computing the *effective exchange rate**, weighting bilateral exchange rates with the share of each other country

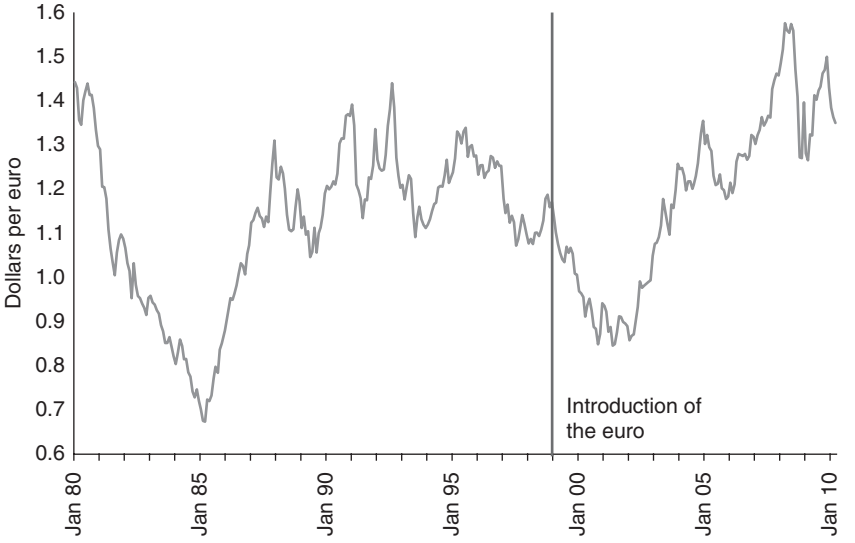


Figure 5.7 The euro–dollar nominal exchange rate, 1980–2010.
Source: Banque de France and Reuters.

Note: Beginning of each month; synthetic euro before 1999.

as a trading partner. Both concepts (real/nominal, effective/bilateral) can be combined (box 5.1).

Box 5.1 Different Measures of Exchange Rates

Let S_{ij} be the nominal exchange rate between currency i and currency j, expressed as the price of currency i in units of currency j. P_i is the price index in country i and P_j the price index in country j. The bilateral real exchange rate is:

$$Q_{ij} = S_{ij}\frac{P_i}{P_j} \tag{B5.1.1}$$

A rise of Q_{ij} is called a real appreciation of currency i and reflects a rise in the relative price of country i vis-à-vis country j. Conversely, country j experiences a real depreciation vis-à-vis country i.

Now let α_j be the share of country j in the foreign trade of country i, with $\sum_j \alpha_i = 1$. One can define the nominal effective exchange rate of country i as a weighted average of its bilateral exchange rates with its trade partners and write it as:

$$E_i = \prod_j S_{ij}^{\alpha_j} \tag{B5.1.2}$$

The *real effective exchange rate** of country *i* is:

$$Q_i = \prod_j Q_j^{\alpha_j} \qquad \text{(B5.1.3)}$$

A rise of Q_i reflects a rise in the relative price of country *i* on average vis-à-vis its trading partners.

Depending on the purpose of the calculation, a different price index needs to be used. Price competitiveness can be gauged by deflating the nominal exchange rate with producer prices or export prices. Cost competitiveness can be measured using unit labor costs. Terms of trade are calculated by comparing export prices with import prices. Finally, a measure of the relative purchasing power of one currency unit is obtained through deflating the nominal exchange rate with consumer price indices. Results can differ widely depending on the price index, as can be seen in figure B5.1.1 showing the real exchange rate of Spain vis-à-vis Germany since the introduction of the euro. Using consumer prices indices as a reference, the real exchange rate of Spain against Germany appreciated by 15% over the period 1999–2008. Using unit labor costs (i.e., the labor cost per unit produced), it appreciated by more than 35%. This suggests that one euro lost more purchasing power in Spain than in Germany during this period, but more strikingly, that Spain lost cost competitiveness against Germany over that period.

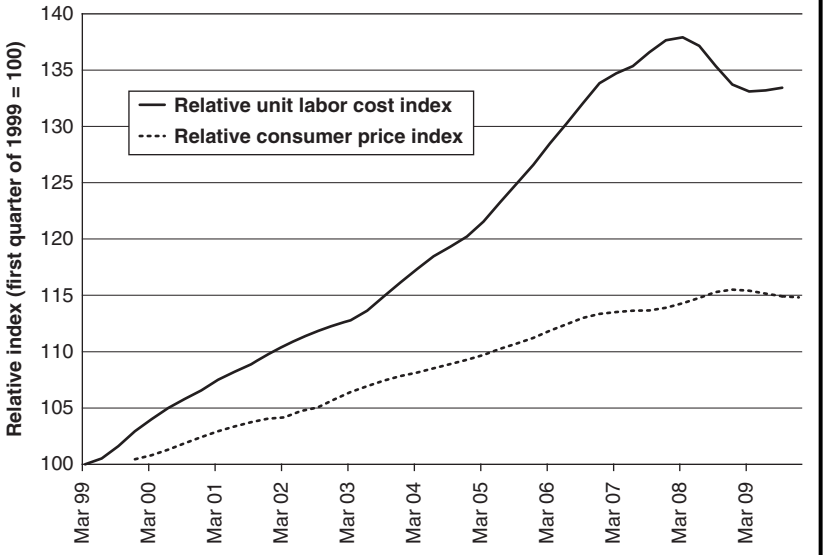


Figure B5.1.1 Purchasing power and cost-competitiveness differential between Germany and Spain, 1999–2009.
Source: OECD, Key Economic Indicators.

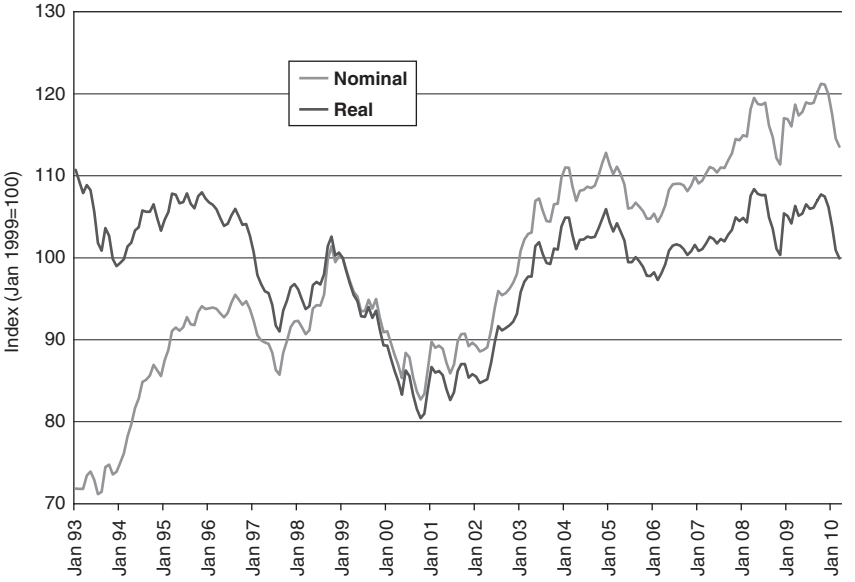


Figure 5.8 Nominal and real effective exchange rates of the euro, 1993–2010.
Source: European Central Bank.

Note: Real effective exchange rate deflated by consumer prices, computed against 52 economies with double export weights.

Observing real exchange rates over long periods and across countries helps uncover two stylized facts.

The first stylized fact relates to the correlation between nominal and real exchange rates: When inflation is low, nominal exchange-rate movements dominate price movements, implying that the real exchange rate is strongly correlated with the nominal exchange rate; when inflation is high, on the contrary, the nominal exchange rate evolves in line with prices and the real exchange rate is relatively stable. These two configurations are illustrated in figures 5.8 and figure 5.9. Figure 5.8 shows the nominal and real effective exchange rates of the euro evolving in close parallel in the 2000s, a period with low inflation in the Eurozone.

In contrast, figure 5.9 shows that, in Brazil from September 1991 to June 1994, the nominal exchange rate was divided by 4000 while prices were multiplied by 5000. With US prices increasing by around 3% a year, the real exchange rate of Brazil depreciated by only 28% during this period, as opposed to 99975% for the nominal exchange rate.

The second stylized fact relates to the long-term behavior of the real exchange rate: In advanced economies, the real exchange rate tends to revert to a constant value in the long run, while in developing countries, the real

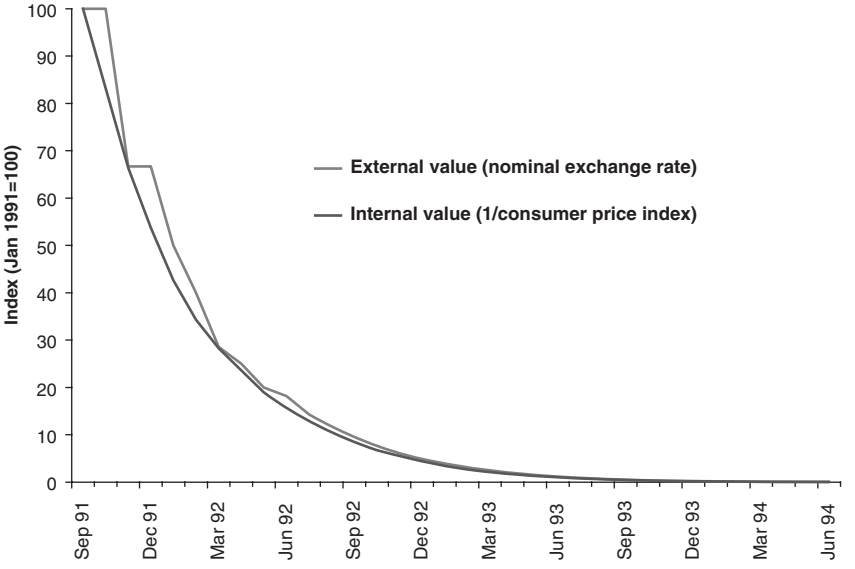


Figure 5.9 Internal and external value of the Brazilian real, 1991–1994.
Source: International Monetary Fund.

exchange rate is not stable in the long run and appreciates as the country develops. The link between the real exchange rate and GDP per capita is robust, as can be seen in figure 5.10. It is called the Balassa–Samuelson effect and we shall explain it later in this chapter.

5.2 Theories

Both the need for and feasibility of an exchange-rate policy are debated among economists. Indeed, the exchange rate can be viewed as a relative price that should be allowed to adjust freely to keep the foreign-exchange market, and through it the economy, at equilibrium. For instance, a negative shock on foreign demand reduces the *current-account surplus* and the demand addressed to domestic firms. The depreciation of the domestic currency (due to the lower demand expressed for it) will stabilize the demand for goods and services (see chapter 4). If the exchange rate does not depreciate by itself following the lower current account, it can do so in reaction to an interest rate cut decided by the central bank to stabilize domestic output.

So why should governments care about the exchange rate? This is because like all asset prices, it is subject to microeconomic shocks: Sales and purchases of currency unrelated to the macroeconomic situation (such as the financing of large-size foreign direct investments), shifts of attitude toward risk, and

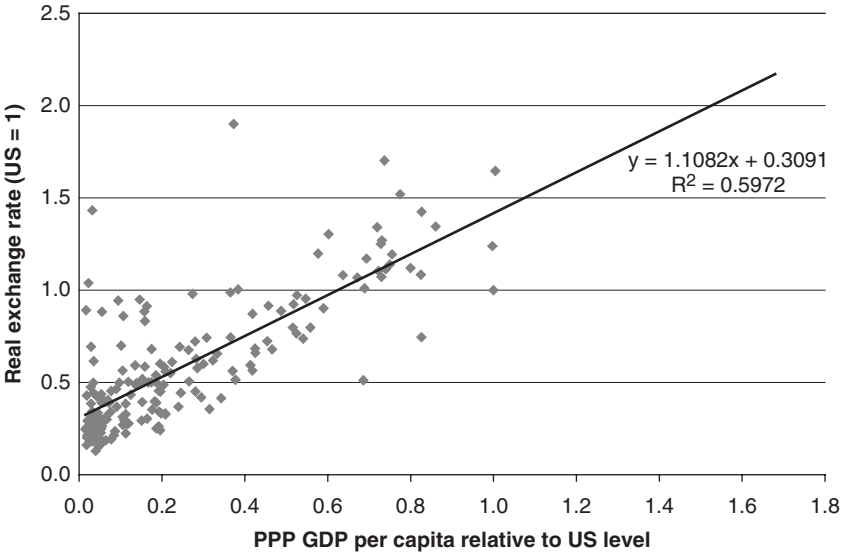


Figure 5.10 PPP GDP per capita and real exchange rates in 2006.
Source: IMF, *World Economic Outlook*, September 2006.

herding behavior of market operators.[21] And because economic agents do not all have the same expectations about the future course of the economy. This has the potential to generate large and persistent deviations between the exchange rate and its economic fundamentals. To the extent that (i) they can be characterized, and (ii) they are costly for the economy (for instance due to hysteresis effects), such deviations need to be opposed. A second question then emerges: That of the policy tools.

Economists are divided on the capacity of a government to manage the exchange rate in a world with free capital movements. In chapter 4, it was shown how the exchange rate can be made to depreciate through a cut in the interest rate. In this chapter, we shall see how foreign-exchange interventions can also be used. However, if interest-rate management is devoted to an internal objective (say price stability), and if there are no obstacles to capital mobility, then the government will fall short of instruments to implement an exchange-rate policy. It may then choose to abandon its monetary independence in order to keep a stable exchange rate, or just drop any exchange-rate objective.

This choice amounts to selecting an exchange-rate regime. It calls for mobilizing various strands of economic theory, ranging from the

21. We will not describe here the vast literature on the microstructure of the foreign exchange market—see Lyons (2001), nor that on the formation of expectations on this market—see De Grauwe (2000).

Mundell–Fleming model to the literature on optimum currency areas, not to mention the more recent theories of currency crises.

5.2.1 Equilibrium exchange rates

As already mentioned, the value of the exchange rate is led by the balance of payments, which describes all transactions concurring to supply and demand of foreign currency against domestic currency. Shocks to the exchange rate originate either in the current transactions account (i.e., in goods and services markets) or in the financial account (i.e., in financial markets). Until the early 1990s, capital movements were restrained in most countries and the former type of shocks was dominant. Since the 1990s, capital movements have become an order of magnitude higher than goods and services transactions, and the latter type of shock has gained importance. Such evolution has paved the way to greater disconnection between exchange-rate variations and the needs of the "real" economy. For instance, the US dollar failed to adjust downward in the early 2000s despite growing current-account deficits in the US. Policymakers need analytical tools to disentangle exchange-rate variations that correspond to the evolution of economic fundamentals, such as the current account or productivity, from those that consist of short- and medium-run deviations from a long-run norm.

a) Purchasing power parity and the Balassa–Samuelson effect

In the long run, there is no reason why, when converted into the same currency at market exchange rates, the level of prices should differ across economically integrated countries. Indeed, when a good is tradable, its price should equalize across countries by virtue of the *law of one price**. If some price differentials do survive, this must be due to transportation costs, tariffs, and other trade barriers, or market imperfections such as imperfect information or monopolistic power. However, if price differentials are structural, they are expected to remain stable or to evolve slowly. Even when they do not, the real exchange rate should revert to a stable level at the aggregate level, based on a macroeconomic adjustment mechanism already identified by David Hume in the seventeenth century as the *price-specie flow mechanism**, a self-stabilizing property of the balance of payments. According to Hume, a country enjoying an increase in price competitiveness normally experiences an improvement in its current account. In a fixed nominal exchange-rate regime, it accumulates foreign-exchange (or gold) reserves. If not sterilized, reserve accumulation inflates the money supply and therefore the general level of prices. The real exchange rate thus appreciates until it is back to its initial level. The price-specie mechanism framed the functioning of the Gold Standard and led to the recognition of relative price variations as a macroeconomic adjustment mechanism (Cassel, 1921). A case in point is the behavior of Europe and the US during World War I. The US economy was stimulated by military

expenses while France and Britain were plagued by war. Under the Gold Standard, a fixed nominal exchange-rate regime, the French franc and the pound sterling could not depreciate against the dollar. The only way to engineer real exchange-rate depreciation was a fall in the prices of British and French goods relative to US goods. In a flexible exchange-rate regime, the current-account surplus in the US would have caused a nominal appreciation of the US dollar, bringing about the same result in real terms.

Price equalization across two countries is called *absolute purchasing-power parity** whereas the stability of price differentials is labeled *relative purchasing-power parity**. The *purchasing power parity exchange rate** is the nominal exchange rate equalizing prices across two countries. With relatively low inflation, the relative price level moves smoothly over time; hence, the PPP level of the exchange rate moves smoothly and offers a benchmark for the observed exchange rate, which is said to be *overvalued** in terms of PPP if it is stronger than the PPP level and *undervalued** in the opposite case.

In the 1980s, economists would usually argue that PPP does not hold even in relative terms or even in the long run. This conclusion was based on time-series analyses of key exchange rates over the 1970s and 1980s. Since the 1990s, longer-time and higher-frequency series, together with the use of panel-data analysis, (both of which involve an increase in the number of observations included in the regressions), have led to a different conclusion. It has increasingly been recognized that there is some mean-reversion toward a stable real exchange rate among the most advanced economies, although the convergence is very slow: On average, it takes three to five years to close half of the gap between the real exchange rate and its long-term value (Rogoff, 1996), which means that if the exchange rate is overvalued by 10% in one given year, it will still be overvalued by 5% after 3–5 years, absent new shocks. Hence, large and persistent fluctuations in the real exchange rate, as evidenced, e.g., in figure 5.8, are not inconsistent with a slow reversion toward PPP.

The problem with the law of one price is that many sectors are shielded from international competition. Hairdressers and restaurants are famous examples, the retail trade and public services are others. In these sectors, the law of one price cannot be expected to hold unless cost and market structures are the same and preferences are identical. The magazine *The Economist* publishes a PPP real exchange-rate index based on the price of hamburgers, the "Big Mac index." In March 2010, the price of an internationally standardized hamburger, the Big Mac, ranked from 1.83 equivalent US dollars in China, to 3.58 dollars in the US and 6.87 dollars in Norway. The law of one price would imply a 49% under-valuation of the Chinese currency, the renminbi, and a 91% over-valuation of the Norwegian krone with respect to the dollar. However, this should not be expected to apply: Hamburgers are not shipped abroad and have a high content of nontradable services that do not bear the same cost across countries.

More generally, the price of nontradable goods tends to be higher in more developed economies. Among the most advanced countries, these differences

are limited and PPP still holds in the long run. However, price differentials in trade-sheltered sectors are very substantial between countries of different development levels, leading PPP to be invalidated.

In 1964, in separate contributions, Bèla Balassa and Paul Samuelson highlighted the role of productivity differentials in explaining such differences in price levels. Their idea goes as follows. Nontradable goods are produced more or less in the same way in all countries, developed or not. In short, a hairdresser has more or less the same number of clients per day in all countries. Productivity in the trade-sheltered sectors is thus comparable.[22] In sectors exposed to international competition, in contrast, productivity is much higher in developed countries than in developing countries, thanks to technical progress. In developing countries, wages have therefore to be lower, so that prices can be the same and abide by the law of one price. However, since workers can move across industries, low wages in the exposed sector will exert downward pressure on wages in the sheltered sector. Finally, the sheltered sector enjoys lower wages than in developed countries, even though productivity is comparable: The price of nontraded goods can thus be lower. Coming back to the previous example, the price of a haircut is *much* lower in developing countries. This explains why the aggregate price level is lower in developing countries.

Now, how do things evolve over time? As productivity in the exposed sector converges toward the level in developed countries, wages increase gradually. In the sheltered sector, however, wage increases are not compensated by a better productivity and have to be passed to consumers. The aggregate price index therefore increases over time compared to that of advanced economies, implying a real exchange-rate appreciation. This is the *Balassa–Samuelson effect** (the math is developed in box 5.2).[23]

This catch-up in prices in developing countries is a natural process which should not be opposed. It does not mean in any respect that competitiveness deteriorates. It has important policy consequences. For instance, it implies that inflation rates are likely to remain quite dispersed in the euro area, creating a challenge for macroeconomic stability as they make real interest rates differ across countries, until GDP per capita has substantially converged across the member countries. Upstream, it means that there is a contradiction in the convergence criteria European countries have to abide by in order to be allowed in the euro area. According to the Maastricht Treaty, candidate countries are required both to achieve nominal exchange rate stability vis-à-vis the euro and to achieve a rate of inflation close to the inflation rate of the best-performing euro area countries. This is incompatible with the

22. This is only partly true since technical progress also enhances productivity in sheltered sectors. US and European shopkeepers benefit from highly advanced payment and supply-chain management schemes, which is not the case for most Chinese shopkeepers.

23. Note that in the very long run, when productivity is equalized internationally, the price of both traded and nontraded goods should equalize, and PPP should apply across all countries.

Balassa–Samuelson effect, since it amounts to blocking real exchange-rate appreciation.[24]

Even more than PPP, the Balassa–Samuelson benchmark is a long-run one since it relies on economic catching up. Assuming that each year 2% of the gap in GDP per capita relative to the leading economy is closed, it takes 30 years for half of the undervaluation relative to PPP to be closed. Obviously, exchange-rate policy cannot rely on this single, very-long-run benchmark.

Box 5.2 The Balassa–Samuelson Effect

Consider a small country with two industries: A traded-goods sector T, say manufacturing, with weight $0 < \alpha < 1$, and a nontradable goods sector N, say services, with weight $1 - \alpha$. The law of one price holds only in the traded-goods sector:

$$Q^T = \frac{SP^T}{P^{T*}} = 1 \qquad \text{(B5.2.1)}$$

where Q^T is the relative price of traded goods between the home country and the foreign one, S is the nominal exchange rate (value of one unit of domestic currency in terms of foreign currency) and P^T, P^{T*} are the price levels of traded goods at home and abroad, respectively. The productivity of the traded-goods sector a^T differs from country to country while the productivity of the nontradables sector a^N is identical: $a^T \neq a^{T*}$, $a^N = a^{N*}$.

Suppose that the nominal wage W is the same throughout a given country due to workers' mobility. Profit maximization under perfect competition leads to a real wage equal to productivity so that:

$$P^N = \frac{W}{a^N}, \quad \text{with } W = a^T P^T$$
$$P^{N*} = \frac{W^*}{a^{N*}}, \quad \text{with } W^* = a^{T*} P^{T*} \qquad \text{(B5.2.2)}$$

Noting that $a^N = a^{N*}$ and that $SP^T = P^{T*}$, one obtains the relative price of nontradable goods:

$$Q^N = \frac{SP^N}{P^{N*}} = \frac{SW}{W^*} \times \frac{a^{N*}}{a^N} = \frac{a^T}{a^{T*}} \qquad \text{(B5.2.3)}$$

24. Assuming that the nontraded goods sector represents half of the economy and that the productivity of the exposed sector catches up with the level of the euro area at a rate of 4% a year, the Balassa–Samuelson effect implies a real exchange-rate appreciation of 2% a year. If the nominal exchange rate is held constant, this means that inflation is higher by 2% than the euro area average. With an average eurozone inflation of 2%, one obtains an inflation rate of 4%. But the Maastricht Treaty's inflation-convergence criterion requires inflation not exceeding the average of the three least-inflationary countries by more than 1.5 percentage points. If this reference level is 1%, then inflation in candidate countries cannot exceed 2.5% for them to join the euro.

Let the aggregate price index P be the weighted average of price indices in both sectors:

$$P = \left(P^T\right)^{\alpha} \left(P^N\right)^{1-\alpha} \qquad P^* = \left(P^{T^*}\right)^{\alpha} \left(P^{N^*}\right)^{1-\alpha} \tag{B5.2.4}$$

Combining equations B5.2.1, B5.2.3 and B5.2.4 finally yields the real exchange rate:

$$Q = \frac{SP}{P^*} = \left(\frac{a^T}{a^{T^*}}\right)^{1-\alpha} \tag{B5.2.5}$$

The real exchange rate depends directly on the productivity differential in the traded-goods sector. If the productivity of a country grows more quickly than that of its trading partners, then its real exchange rate appreciates. Depending on the exchange-rate regime, this appreciation can materialize through nominal exchange-rate appreciation or through higher inflation.

b) Real exchange rates and the balance of payments

In the shorter term, the real exchange rate may depart from PPP or from the Balassa–Samuelson benchmark due to macroeconomic imbalances, especially balance-of-payment imbalances. This requires other approaches to the equilibrium exchange rate that link real exchange-rate developments to internal and external imbalances.

The exchange rate determines the relative price of goods and services produced in the domestic economy, hence domestic demand for foreign goods, foreign demand for domestic goods, and the price of imported goods in the consumption basket.

Denoting C private consumption, I private investment, G public expenditures, X exports and IM imports, all expressed in units of domestic output, aggregate demand Y^d is written as:

$$Y^d = C + I + G + X - IM \tag{5.1}$$

Denoting Q the relative price of domestic goods in terms of foreign goods (i.e., the real exchange rate of the domestic country), we have $IM = M/Q$, where M is the volume of imports expressed in units of foreign goods and services. The accounting equation can be rewritten accordingly:

$$Y^d = C + I + G + B \tag{5.2}$$

where B is the trade balance expressed in domestic output units:

$$B = X - \frac{M}{Q} \tag{5.3}$$

A real exchange-rate depreciation (a decrease in Q) has three distinct effects on the trade balance B, and hence on aggregate demand Y^d: (i) It increases the volume of exports X due to higher price-competitiveness; (ii) it reduces the volume of imports M because imported goods and services are more expensive; (iii) it raises the relative value of each imported unit. The first two are volume effects and they influence positively the trade balance B; the last one is a valuation effect which influences it negatively. The net effect of the real exchange-rate depreciation on B is thus ambiguous: B rises if the volume effects dominate the valuation effect. This condition is called the *Marshall–Lerner condition** (see box 5.3).

Box 5.3 The Marshall–Lerner Condition and the J-curve

The Marshall–Lerner condition sets the condition for a currency depreciation to impact positively on the trade balance. Denote by X,M, and B exports, imports, and the trade balance respectively. X and B are expressed in units of domestic production, while imports M are expressed in units of foreign production. To express all flows in units of national production, imports must be divided by the real exchange rate Q. Hence the trade balance in units of national goods is written as:

$$B(Q) = X(Q) - M(Q)/Q \tag{B5.3.1}$$

A real depreciation (fall in Q) leads to a rise in X and a fall in M. Both volume effects have a positive impact on the trade balance B. Now, a real depreciation also revalues every imported good in terms of domestic goods. This price effect has a negative impact on B. The net impact of the real exchange-rate depreciation on B is positive if volume effects dominate the price effect. Denoting by ρ the initial coverage ratio of imports by exports ($r = QX/M$), we have:

$$\frac{dB}{X} = \frac{dX}{X} - \frac{1}{\rho}\frac{dM}{M} + \frac{1}{\rho}\frac{dQ}{Q} \tag{B5.3.2}$$

Assuming exogenous income in both the domestic economy and in the rest of the world, and denoting by ε_X and ε_M the price elasticities of exports and imports, respectively ($\varepsilon_X, \varepsilon_M > 0$), we have $dX/X = -\varepsilon_X dQ/Q$ and $dM/M = e_M dQ/Q$. Plugging this into (B5.3.2), we get the variation in B following a variation in Q:

$$\frac{dB}{X} = -\left(\varepsilon_X + \frac{\varepsilon_M}{\rho} - \frac{1}{\rho}\right)\frac{dQ}{Q} \tag{B5.3.3}$$

Assume that $QX = M$ initially ($\rho = 1$). The above relation shows that a depreciation in the real exchange rate ($dQ/Q < 0$) raises the trade balance if and only if:

$$\varepsilon_X + \varepsilon_M > 1 \tag{B5.3.4}$$

i.e., if the reaction of trade volumes dominates the revaluation of imported goods. This condition is called the Marshall–Lerner condition (see Lerner, 1944).

If the trade balance is in surplus initially ($\rho > 1$), then the Marshall condition is more easily met because the revaluation effect ($1/\rho$) is relatively smaller. The reverse applies if the trade balance is initially in deficit ($\rho < 1$).

Whereas the revaluation of the imported goods is immediate, the effect of the real exchange rate on the exported and imported volumes can be slow due to information asymmetries and to long-run contracting. This translates into price elasticities that are lower in the short run than in the long run. Consequently, a real depreciation results in a fall in the trade balance in the short run because the price effect dominates the volume effects; then, the exported and imported quantities adjust and the trade balance increases. This differentiated reaction of the trade balance over time is called the *J*-curve.

The Marshall–Lerner condition is generally met in the medium term, i.e., after a few quarters. However, as volumes tend to react slowly to relative price variations while the valuation effect is immediate, an exchange-rate depreciation triggers an immediate deterioration in the trade balance in the short run before the trade balance improves. Hence the reaction of the trade balance following an exchange-rate depreciation is *J*-shaped: This is the *J-curve** (see figure 5.11).

In what follows, the Marshall–Lerner condition is assumed to hold. Aggregated demand Y^d is thus a decreasing function of the real exchange rate Q, and therefore a decreasing function of both the domestic price index P and of the nominal exchange rate S rate.[25] In figure 5.12, aggregate demand is downward-sloping (like in chapter 1), and it moves upward when the nominal exchange rate depreciates.

We now turn to the impact of the exchange rate on aggregate supply, i.e., on the total volume of goods and services supplied by companies at a given price level P or, equivalently, on the price set by the firms for a given level of supply. In the long term, the neutrality of money entails that aggregate supply is vertical, i.e., aggregate supply does not depend on the aggregate price level. In the short run, however, to the extent that there are nominal rigidities, the aggregate supply is upward-sloping.

In the very short run, a nominal depreciation of the domestic currency moves the demand curve upward while the supply curve is little affected,

25. The impact of a depreciation of the exchange rate on aggregate demand can, however, be negative in a strongly indebted country in foreign currency, because the depreciation revalues the net debt, causing a negative wealth effect.

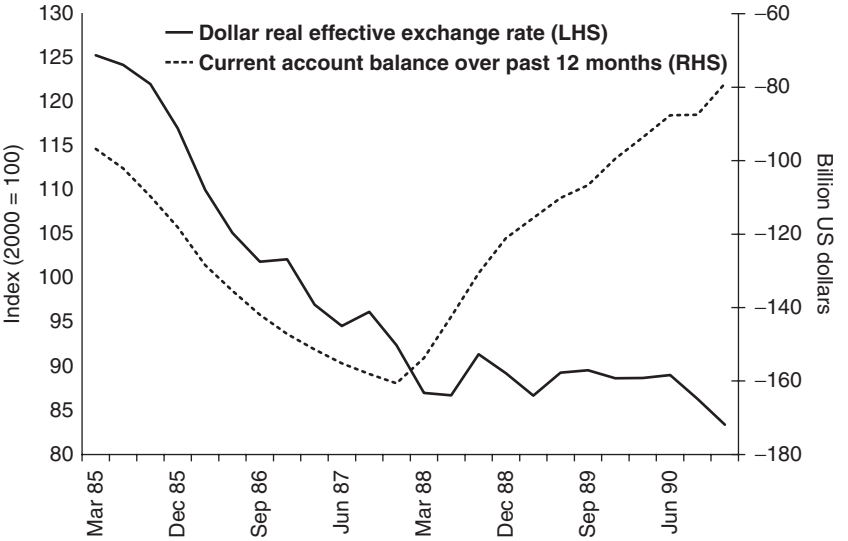


Figure 5.11 The *J*-curve in the US, 1985–90.
Source: Ecowin.

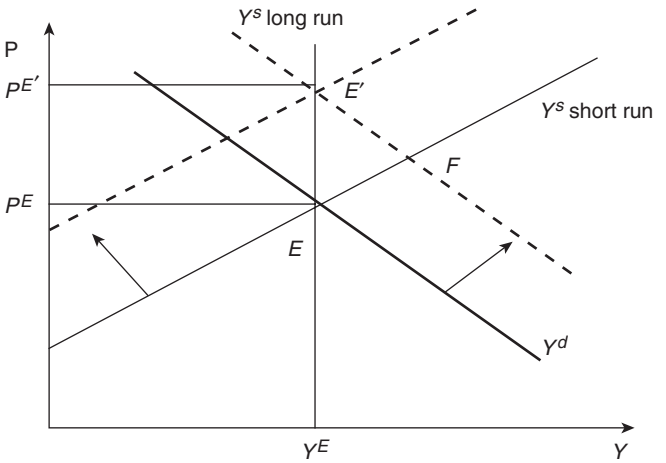


Figure 5.12 Impact of a nominal depreciation on production and prices.
Reading: A nominal depreciation moves the demand curve Y^d upwards. In the absence of second-round impact on inflation, the macroeconomic equilibrium moves from E in F. Then, progressively imported inflation moves the supply curve Y^d upwards. In the long run, output is back to its initial level in E' and the general price index is higher (the real exchange rate is back to its initial level).

due to nominal rigidities. Hence output increases from point E to point F in figure 5.12. Then, firms tend to raise their prices due to more expensive imported inputs and to the necessity to raise nominal wages in order to compensate the workers for losses in purchasing power parity due to the depreciation: The supply curve gradually moves upward. This eventually brings real output back to its initial level while the general level of prices rises. In the long run (E'), output is back to its initial level and all prices have risen in the same proportion as the exchange-rate depreciation.

What complicates the analysis is that firms may be able to discriminate prices across markets: If the nominal exchange rate depreciates, they may not fully pass this depreciation onto export prices and enjoy higher margins. The extent of such *pricing to market** crucially depends on local market structures (see box 5.4). This phenomenon, which goes back to Joan Robinson (1947) and was formally introduced by Paul Krugman (1987), affects the reaction of relative prices, hence of the trade balance, to nominal exchange-rate variations.

Box 5.4 Pricing to Market and Exchange-Rate Pass-Through

The seminal pricing-to-market model assumes monopolistic competition and aims at explaining how firms' mark-ups vary as a function of the exchange rate and under which condition it is optimal for them not to pass the whole exchange-rate variation onto export prices.

Consider a continuum of varieties of the same good $i \in [0, 1]$. The utility of the representative consumer is defined as:

$$U = \left(\int_0^1 c(i)^{\frac{\eta-1}{\eta}} di \right)^{\frac{\eta}{\eta-1}} \qquad \text{(B5.4.1)}$$

where $c(i)$ denotes consumption of variety i and $\eta > 1$ is the elasticity of substitution between varieties. This assumption of a *preference for diversity* is discussed in chapter 6. The budget constraint of the consumer is written as:

$$PC = \int_0^1 p(i)c(i)di = R \qquad \text{(B5.4.2)}$$

where C represents the aggregate consumption volume of the representative household, $p(i)$ denotes the price of variety i, R is the consumer's nominal income and P is the aggregate price index. Utility maximization under budget constraint (B5.4.2) leads to an inverse relationship between the consumption of variety i and its relative price:

$$c(i) = \left(\frac{p(i)}{P} \right)^{-\eta} C \qquad \text{(B5.4.3)}$$

Assume that each variety is produced with labor only, through the following, very simple production function:

$$y(i) = l(i) \tag{B5.4.4}$$

where $y(i)$ denotes the production of variety i and $l(i)$ is employment needed to produce $y(i)$. The firm producing i faces a domestic demand $c(i)$ and a foreign demand $x(i)$, the latter following the same pattern as domestic demand, except for its price which is $q(i)$ on the foreign market. Denoting by e the nominal exchange rate, the export price of i is $q(i)/e$ in the exporter's currency and $q(i)$ in the importer's currency. Hence, the foreign demand is written as:

$$x(i) = \left(\frac{q(i)}{p^*}\right)^{-\eta^*} C^* \tag{B5.4.5}$$

where $\eta^* > 0$ is the elasticity of substitution on the foreign market and C^* denotes total consumption of the foreign representative consumer. Assuming that the nominal wage W is given for the firm, profit maximization yields optimal pricing:

$$\begin{aligned} p(i) &= \frac{\eta}{\eta - 1} W \\ \frac{q(i)}{e} &= \frac{\eta^*}{\eta^* - 1} W \end{aligned} \tag{B5.4.6}$$

The price of i is the same on the foreign as on the domestic market if the elasticity of substitution is the same on both markets. In contrast, if $\eta^* > \eta$, then the price of the same variety is lower on the foreign market than on the domestic one. Denoting by μ and μ^*, respectively, the mark-up on the domestic and on the foreign market, we have: $\mu = \eta/(\eta - 1)$ and $\mu^* = \eta^*/(\eta^* - 1)$. If η and η^* are increasing function of prices, then m and m^* are decreasing functions of prices: The higher prices are, the more sales tend to decline when the price rises further. In this case, the mark-up declines whenever producer costs go up: The cost increase is less than proportionally passed on prices. Costs include domestic wages, but also the exchange rate. Such incomplete pass-through reduces the impact of exchange-rate variations on volumes exported. A larger variation of the exchange rate is then necessary to rebalance the trade account by a given amount.

Empirically, Gaulier et al. (2008) find a very different pass-through coefficient across countries and products, with a relatively high median (82%).

On the whole, however, a nominal depreciation of the national currency results in inflation. The impact on real output is theoretically ambiguous, but in general positive in the short term. The distribution of the adjustment between prices and volumes depends on three factors:

- The sensitivity of trade volumes to price-competitiveness (the extent of the shift of the demand curve).
- The flexibility of supply (the slope of the short-run supply curve): A flexible supply means that the firms are willing to provide more output without a large increase in output prices (flat supply curve); this may appear especially if they have output capacities that are unemployed.
- Imported inflation (the extent of the shift in the short-run supply curve): In a very open economy, or in a partially dollarized economy (in which prices are expressed in foreign currency), an exchange-rate depreciation quickly translates into higher prices. Even in the short run, the depreciation has little impact on real output but a strong impact on prices. Consistently, very open or partially dollarized countries are tempted to fix their exchange rate, or even to give up monetary sovereignty.

c) Equilibrium exchange rates

The above analysis directly yields a concept of equilibrium exchange rate: Assuming that the trade balance reacts positively to a depreciation in the real exchange rate (i.e., that the Marshall–Lerner condition applies), it is possible, based on estimated price-elasticities of exports and imports, to calculate the real exchange rate that would be consistent with a certain level of the trade balance.

Assume, for instance, that a 10% depreciation of the real exchange rate raises the current account by 1% of GDP. Then, a 30% depreciation is needed to bring the US current-account deficit from 5% of GDP (its 2008 level) to, say, 2% of GDP. If inflation is similar in all countries, this means that the nominal exchange rate needs to depreciate by 30%, otherwise this depreciation could also be achieved through lower inflation in the US.

This simple calculation raises the difficult question of the current-account level to be targeted. As in the case with public finances (see chapter 3), the question is less the level of the deficit than the evolution of the debt and its sustainability. Indeed, the *raison d'être* of financial liberalization is to allow countries with excess domestic savings to invest in countries where saving is insufficient. Freedom of capital movements removes the financial constraint on developing countries and allows them to borrow from abroad to finance the catch-up of their capital stock. This involves current-account imbalances that are sustainable if they yield higher GDP and exports in the future.

Accordingly, John Williamson's *fundamental-equilibrium exchange rate** or *FEER** is the real exchange rate "which is expected to generate a current-account surplus or deficit equal to the underlying capital flow over

Table 5.3
FEER estimates of Chinese renminbi undervaluation

Reference	Year under review	Target level of Chinese current account surplus (as % of GDP)	Misalignment against US dollar (–: undervaluation)
Coudert and Couharde (2005)	2003	–1.5%	–44%
Coudert and Couharde (2005)	2003	–2.8%	–54%
Goldstein (2004)	2003	–1%	–15 to –30%
Jeong and Mazier (2003)	2000	–1.5%	–60%

the cycle, given that the country is pursuing 'internal balance' as best as it can and not restricting trade for balance of payments reasons" (Williamson, 1983). The FEER is thus the real exchange rate that achieves both full employment and a "sustainable" current account level. The methodology basically consists in inverting a current-account equation, as explained above (partial-equilibrium approach). A more comprehensive approach uses a macroeconometric model where all macroeconomic variables (and especially the level of output) are endogenous to the exchange-rate adjustment.

Table 5.3 compares various empirical calculations of the FEER of the Chinese currency in 2003, using partial-equilibrium approaches. The estimates differ widely and the only robust conclusion is that a substantial appreciation of the Chinese renminbi was warranted to curb the Chinese current surplus. One major source of fragility of the FEERs is the difficulty of estimating the price elasticities of exports and imports. A second fragility is the difficulty in setting current-account targets, which necessarily involves some normative judgment. Finally, the concept of internal equilibrium may be difficult to implement in a country like China where there is structural excess labor supply in agriculture.

In order to avoid some of these difficulties, Jerome Stein (1994) has defined the natural real exchange rate (NATREX) in the same way as the FEER, but he has further assumed the target current account to be equal to the *ex ante* savings–investment balance, based on fundamentals such as productivity and the rate of time preference (proxied by the ratio of private and public consumption to GDP). For instance, a rise in government consumption leads the fundamental level of savings to fall. The NATREX appreciates in order to reduce excess demand. In the longer term, the accumulation of current-account deficits will require a depreciation of the exchange rate in order for the debt to stabilize. Hence, the NATREX can be viewed as a dynamic version of the FEER.

d) The intertemporal approach of the balance of payments

One step further, why not envision the equilibrium exchange rate within a fully fledged intertemporal, utility-maximizing framework? The *intertemporal*

*approach of the balance of payments** (Obstfeld and Rogoff, 1999) provides such a framework, considering external indebtness in the same way as personal or corporate debt, as a result of rational microeconomic choices. The Mundell–Fleming model of chapter 3 was based on Keynesian premises, with aggregate demand being modeled in an ad-hoc way. In contrast, the intertemporal approach relates external saving and investment to explicit intertemporal utility optimization. The external position of a country is thus the outcome of individual decisions concerning consumption and labor supply. This line of research, which developed in the 1990s, is the backbone of a broader research program aiming at providing international economics with microeconomic foundations, the *new open economy macroeconomics** (see Lane, 2000, for a survey).

Intertemporal models of the balance of payments combine some kind of nominal rigidity in the short run (otherwise the exchange rate would not affect aggregate demand) with price flexibility in the long run (Obstfeld and Rogoff, 1995; Corsetti and Pesenti, 2001). By contrasting the impact of temporary and permanent shocks, they provide an adequate framework for assessing current account sustainability and its implications in terms of exchange rates. For instance, a positive, transitory shock on productivity leads to a current-account surplus and a transitory depreciation of the real exchange rate. This is because households save a larger share of their current income, since they know that productivity and income will fall back in the future. The relative price of domestic goods declines in order for foreign demand to substitute for the absent domestic demand. In the longer run, the relative price of domestic goods adjusts upward; domestic households reallocate their consumption basket in favor of goods produced abroad, and the current-account surplus disappears.

In contrast, the current account reacts negatively to a positive, permanent shock on productivity because households immediately raise their consumption level, knowing that their income will be higher in the future. Consistent with this, it has sometimes been argued in the late 1990s to early 2000s that the large current-account deficit in the US was sustainable because it was the rational reaction of US households to the permanent productivity shock brought about by the new information technologies. However, this explanation has become less potent in the 2000s with the persistence of the deficit and the subsequent global financial crisis.

Box 5.5 Exchange Rate and Intertemporal Adjustment

Consider a small open economy where households live in two periods. At each period, they consume a tradable good T which can either be produced locally or imported, and a nontradable good N which is produced locally and cannot be imported. The consumption level of the representative household is denoted c_t^i, where t denotes the period ($t = 1, 2$) and i

the good ($i = N, T$). All decisions concerning consumption are taken at period 1 and there is no uncertainty. The utility function of the representative household is assumed to be log-linear:

$$u\left(c_t^N, C_t^T\right) = \gamma \log c_t^N + (1-\gamma)\log c_t^T, \quad \text{with } 0 < \gamma < 1 \quad \text{(B5.5.1)}$$

Let us denote by E_t the relative price of the nontradable good in terms of the tradable one at period t ($E_t = p_t^N / p_t^T$), R_t the household's real income (expressed in terms of the tradable good) at period t, r the real interest rate and β (with $\beta < 1$) the discount factor (which measures preference for present). The representative household's program is then written as:

$$\underset{c_t^i}{Max}\, U = u\left(c_1^N, c_1^T\right) + \beta u\left(c_2^N, c_2^T\right)$$

$$s.c. \quad \left(E_1 c_1^N + c_1^T\right) + \frac{1}{1+r}\left(E_2 c_2^N + c_2^T\right) \leq R_1 + \frac{R_2}{1+r} \quad \text{(B5.5.2)}$$

To solve this equation, one writes the Lagrangian $L = U - \lambda(C - R)$, where C denotes intertemporal consumption, $R = R_1 + \frac{R_2}{1+r}$ is the intertemporal income and λ is the Lagrange multiplier. Then, the partial derivative of L relative to each consumption volume c_t^i is calculated, and the Lagrange multiplier can be eliminated by deriving intra- and intertemporal consumption ratios as follows:

$$\frac{c_2^T}{c_1^T} = \beta(1+r) \qquad \frac{c_1^T}{c_1^N} = \frac{\gamma}{1-\gamma}E_1$$

$$\frac{c_2^N}{c_1^N} = \beta(1+r)\frac{E_1}{E_2} \qquad \frac{c_2^T}{c_2^N} = \frac{\gamma}{1-\gamma}E_2 \quad \text{(B5.5.3)}$$

The left-hand column of (B5.5.3) provides the two conditions for the intertemporal optimization of consumption (optimal allocation of consumption for the same good in two different periods) whereas the right-hand column represents the two conditions for intratemporal optimization (optimal allocation of the consumption of two different goods at the same period). From (B5.5.3), the four consumption levels can be recovered as functions of intertemporal income R:

$$c_1^T = \frac{\gamma}{1+\beta}R \qquad c_1^N = \frac{1-\gamma}{1+\beta}\frac{R}{E_1}$$

$$c_2^T = \frac{\gamma\beta}{1+\beta}(1+r)R \qquad c_2^N = \frac{(1-\gamma)\beta}{1+\beta}(1+r)\frac{R}{E_2} \quad \text{(B5.5.4)}$$

An increase in productivity raises intertemporal income R: Consumption increases for both goods at both periods. Note that

consumption rises at period 1 even if productivity increases only at period 2, and vice versa. If productivity rises only in the tradable sector, then the production of this sector increases, but not that of the nontradable sector whereas consumption increases for both goods. Market equilibrium then requires an increase in the relative price of nontradables E_t. The real exchange rate then appreciates, since it is positively related to E_t: Denoting by Q_t the real exchange rate and by S_t, the nominal one, we have:[a]

$$Q_1 = S_t \frac{\left(p_t^T\right)^\gamma \left(p_t^N\right)^{1-\gamma}}{\left(p_t^{T*}\right)^\gamma \left(p_t^{N*}\right)^{1-\gamma}} = \frac{S_t p_t^T}{p_t^{T*}} \left(\frac{p_t^N / p_t^T}{p_t^{N*} / p_t^{T*}}\right)^{1-\gamma} \tag{B5.5.5}$$

where asterisks refer to foreign prices. Assuming that the law of one price applies to the tradable sector (i.e., $S_t p_t^T = p_t^{T*}$), we get:

$$Q_t = \left(\frac{E_t}{E_t^*}\right)^{1-\gamma} \tag{B5.5.6}$$

Note that, by definition, we have, at each period, $c_N = y_N$: Since N is nontradable, its consumption per inhabitant must equalize its production per inhabitant. From equations (B5.5.4) and (B5.5.6), a relation between the growth rate of nontradable production and that of the real exchange rate can then be derived:

$$\frac{y_2^N}{y_1^N} = \beta(1+r)\frac{E_1}{E_2} = \beta(1+r)\left(\frac{Q_1}{Q_2}\right)^{1/(1-\gamma)}\left(\frac{E_1^*}{E_2^*}\right) \tag{B5.5.7}$$

Absent productivity growth in the nontradable sector, y^N is constant and the variation of the real exchange rate only depends on the evolution of the relative price of nontradables abroad and on the discount factor (compared to the real interest rate). Suppose the relative price of nontradable abroad stays constant ($E_1^* = E_2^*$). If domestic consumers are impatient ($\beta(1+r) < 1$), they intend to consume more of their intertemporal income during the first period. To this purpose, imports of tradables must exceed exports, which amounts to a trade deficit. During the second period, the debt contracted during the first period must be paid back. The real exchange rate appreciates during the first period so as to encourage households to consume more tradables (whose relative price falls); it depreciates in the second period so as to discourage them from consuming tradables (whose relative price then increases).

Similarly, a rise in productivity in the tradable, sector, if expected in period 1 for period 2, produces a trade deficit in period 1and a surplus in period 2. For the same reasons as above, the real exchange rate needs to appreciate in period 1 and depreciate in period 2. The profile of the exchange rate depends on the transitory or permanent pattern of the productivity shock.

The fundamental lesson from the intertemporal approach is that international capital markets allow households to see their budget constraints loosened due to the possibility of lending to or borrowing from abroad, and that real exchange rates, by modifying the relative price of nontradables in terms of tradables, affects the allocation of domestic consumption between both types of goods so as to ensure intertemporal balance. Hence, this results in calculating the real exchange-rate path that is consistent with households' intertemporal budget constraint, i.e., with current-account sustainability (see, e.g., Obstfeld and Rogoff (2007) on the US case).

In the simple version of the model studied here, the price of domestic tradables in terms of foreign ones stays constant (law of one price). This model can however be combined with a monopolistic–competition model where tradables are imperfect substitutes. In this case, productivity shocks affect the real exchange rate both through the relative price of nontradables in terms of tradables, and through the relative price of tradables in terms of foreign tradables.

In all cases, the model focuses on the real exchange rate. When combined with a specification of nominal rigidities, it can describe how the nominal exchange rate carries out most of the adjustment in the short run, although not in the long run.

[a] The real exchange rate here refers to the relative price of the consumption basket compared to the price of the same basket abroad. In turn, E_t represents the relative price of the nontradable good in terms of the tradable one. It is sometime referred to as the *internal* real exchange rate.

Like PPP or the Balassa–Samuelson effect, these different approaches can be used as benchmarks to assess whether a currency is overvalued or undervalued and to provide an order of magnitude of the misalignment. Although they are generally based on direct or indirect econometric estimations (meaning that there is some explanatory power over the estimation period[26]), these benchmarks have low predictive power. Hence they cannot be used, for instance, to assess the impact of policy measures such as a change in short-term interest rates, an official intervention or a fiscal policy. To this aim, a realistic model of exchange-rate determination is needed. Obviously, such a model must account for the behavior of international investors.

26. The intertemporal model of the balance of payments can be estimated, and shocks identified, through dynamic stochastic general equilibrium (DSGE) methodology. In the FEER case, the exchange rate derives from inverting an estimated equation of the current account. Although no exchange-rate equation is estimated, it has been shown that the FEER nevertheless cointegrates with the observed exchange rate (see Barisone et al., 2006).

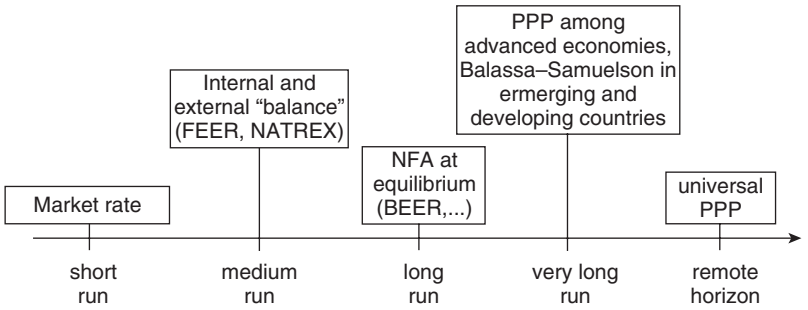


Figure 5.13 Equilibrium exchange rates at various time horizons.
Source: Bénassy-Quéré, Béreau and Mignon (2010).

Figure 5.13 summarizes the various concepts of equilibrium exchange rates, from short-run market equilibrium to the very long run of universal price convergence.

e) The portfolio model and net foreign asset accumulation

In chapter 4, the Dornbusch exchange-rate overshooting model was used to explain the magnifying effect of interest-rate variations on nominal exchange rates. From the inception of the post-Bretton-Woods system, Rudiger Dornbusch had foreseen that floating exchange rates would be much more volatile than their macroeconomic determinants, such as money, interest rates, or prices. In the Dornbusch model (box 4.16), the exchange rate is determined by money supply in the long run. Neither trade nor portfolio flows play a role. This is due to the hypotheses of perfect capital mobility, perfect substitutability between assets, and risk-neutrality that provide the basis for uncovered interest rate parity.

Still, interest rates alone cannot explain exchange-rate variations. The strong depreciation of the euro in 1999–2000 was partly explained by an interest rate differential favoring the US dollar, but massive foreign direct investments from Europe into the US also had a material effect. Likewise, the depreciation of the dollar in 2002–03 reflected the reversal of the interest-rate differential, but also the widening of the US current-account deficit and the drying up of foreign direct-investment inflows. Finally, the stabilization of the dollar in 2004–06 was the result of the contradictory effects of a widening current-account deficit and a higher interest rate in the US than in the euro area.

By relaxing the assumption of risk-neutrality of asset holders, the *portfolio-choice model** (box 5.6) encompasses these various explanations of exchange-rate fluctuations. In this model, a rise in the domestic interest rate, a fall in the foreign interest rate, an increase of the net external position, a decrease in

Table 5.4
Determinants of the euro–dollar exchange rate, 1999–2007

1999	2000	2001	2002	2003	2004	2005	2006	2007
Euro/USD (USD for 1 euro)								
1.07	0.92	0.89	0.95	1.13	1.24	1.24	1.26	1.37
US current account, USD bn								
−299.8	−415.2	−389.0	−472.4	−527.5	−665.3	−791.5	−869.1	−731.3
Euro area current account, euro bn								
−23.9	−88.8	−22.1	57.0	32.4	55.6	18.1	−1.3	26.6
Net FDI flows from the euro area to the US, USD bn								
62.3	126.2	2.2	−41.0	−10.0	−67.0	−44.4	40.2	−21.1
Three-month interest-rate differential (euro–USD)								
−2.5%	−2.1%	+0.5%	+1.2%	+1.1%	+0.5%	−1.4%	−2.1%	−1.0%

Sources: European Central Bank and US Bureau of Economic Analysis.

desired net foreign assets or in foreign-exchange reserves all lead to a nominal appreciation of the domestic currency.

Box 5.6 The Portfolio-Choice Model

The portfolio-choice model (Branson et al., 1977) highlights the role of financial assets in exchange-rate determination. Consider a small, open economy where capital is free to move in and out of the country. The equilibrium of the balance of payments can be written as:

$$B = d(F/S) + d(R/S) \tag{B5.6.1}$$

where S designates the nominal exchange rate, B the current-account surplus in domestic currency, F the net holding of foreign securities by residents, R the foreign-exchange reserves outstanding, and d designates a variation per unit of time. Both F and R are expressed in foreign currency. This simply expresses the equality between the current-account balance and the financial-account balance. Over time, this equality translates into a relationship between stocks of assets:

$$A = (F + R)/S \tag{B5.6.2}$$

where A is the net foreign asset position, i.e., the difference between gross assets and gross liabilities. A can either reflect a net claim on foreign countries ($A > 0$) or a net liability ($A < 0$). It derives from the accumulation of current-account imbalances as well as valuation effects.

Let W be total private sector wealth expressed in real terms. It is made of domestic assets D and foreign assets F, deflated by the aggregate price level P:

$$W = (D + F/S)/P \tag{B5.6.3}$$

The portfolio approach is based on Tobin's work on optimal portfolio allocation (Tobin, 1958). Portfolio allocation stems from an arbitrage between risk and return. At each moment in time, private agents choose the share of their wealth invested in foreign currency, $f = F/SPW$, that maximizes their utility. As an approximation of a more general concave-utility function, utility depends positively on the expected increase in W and negatively on its variance:

$$Max\, U = E\left(\frac{dW}{W}\right) - \frac{\alpha}{2} VAR\left(\frac{dW}{W}\right) \tag{B5.6.4}$$

where E stands for the expected value, VAR for the variance per unit of time, and $\alpha \geq 0$ is the risk-aversion coefficient. With risk originating in the exchange rates and in the price level, it can be shown that the solution is:

$$f = f_0 + \frac{i^* - i - \mu_s}{\alpha \sigma_s^2}, \quad \text{with} \quad f_0 = -\frac{\sigma_{sp}}{\sigma_s^2} \tag{B5.6.5}$$

where i is the domestic interest rate, i^* the foreign interest rate, $\mu = E(dS/S)$ and $\sigma_s^2 = VAR(dS/S)$ are the expected variation and volatility of exchange-rate variations per unit of time, and $\sigma_{sp} = cov(dS/S,\ dP/P)$ is the covariance between inflation and the exchange-rate variations. The share of wealth held in foreign currency is thus the sum of two terms:

- The first term f_0 gives the proportion of foreign assets that would be held by an investor with infinite risk aversion. The corresponding portfolio is called the *minimum-risk portfolio**. If the exchange rate goes down when prices go up ($\sigma_{sp} < 0$) then the investor will invest part of his or her wealth in foreign assets as a hedge against purchasing power losses due to inflation. If the exchange rate moves in line with prices ($\sigma_{sp} > 0$), then he or she will borrow in foreign currency.
- The second term involves the expected yield differential as in the uncovered interest-rate parity. Indeed, uncovered interest parity is a specific case of the portfolio model.

Equations (B5.6.2) and (B5.6.5) yield:

$$i = i^* - \mu_s + 2\alpha\sigma_s^2 \left(f_0 + r - a\right) \tag{B5.6.6}$$

where $r = R/SPW$ and $a = A/PW$. If investors are risk-neutral ($\alpha = 0$), this amounts to the uncovered interest rate parity of chapter 4:

$$i = i^* - \mu_s \tag{B5.6.7}$$

If investors are risk adverse, a *risk premium**, the last term in equation (B5.6.6), is needed to remunerate investors who are willing to depart from their minimum-risk allocation. Another way to understand this equation is to write it as:

$$s = Es + i - i^* + 2\alpha\sigma^2\left(a - f_0 - r\right) \qquad \text{(B5.6.8)}$$

where $s = \ln S$ and $\mu_s = Es - s$. A rise in the domestic interest rate i, a fall in the foreign interest rate i^*, an increase in the net foreign asset position ratio a, a decrease of the minimum-risk net foreign-asset position f_0 or of foreign-exchange reserves r all lead to an appreciation of the domestic currency (s rises).

According to the portfolio model, the continuing deterioration of the net foreign-asset position of the US since the mid-1980s (see figure 5.6) should lead non-US investors to demand higher yields on dollar-denominated assets if they are to continue to hold them; this can be obtained through a fall in the dollar exchange rate that makes US assets cheaper.

The portfolio-choice model also explains the conditions under which central bank or government intervention in the foreign-exchange market can be efficient in changing the exchange rate. As shown in box 5.3, three policy tools can be used: Sale and purchase of foreign-exchange reserves, interest rate changes, and communication aiming at monitoring exchange-rate expectations. The former instrument has little direct impact if risk aversion is low and capital mobility is high. The use of these three instruments to carry out an exchange-rate policy is further discussed in section 5.3.

Turning back to equilibrium exchange rates, the portfolio-choice approach provides a long-run benchmark for the exchange rate, provided short-run factors such as the interest-rate differential are cancelled out. Accordingly, Clark and MacDonald (1998) deduce the *behavioral-equilibrium exchange rate** or *BEER** from a stable long-run relationship between the real exchange rate and its macroeconomic determinants taken from the portfolio-choice approach (mainly the net foreign asset position and the terms of trade, plus the interest-rate differential as a short-run control) and from the Balassa–Samuelson approach (productivity differentials). From an econometric standpoint, this amounts to identifying a co-integrating relation linking these determinants. The BEER is a mere historical regularity and does not ensure in any respect the sustainability of the current account. For instance, the BEER of the US dollar may be estimated for a period where the appetite of world investors for the dollar was very high, e.g., due to the lack of an alternative to the dollar as an international store of value. Hence, this approach is more a complement to than a substitute for the FEER approach.

A key channel of current account adjustment is the *valuation effect**, i.e., the effect of nominal exchange rate changes on the value of assets and liabilities. This is because the year-on-year variation in the net foreign-asset position depends not only on current transactions and capital flows, but also on capital gains or losses on foreign assets (like foreign equities and bonds held by residents) and on foreign liabilities (like domestic equities and bonds held by nonresidents). Valuation effects triggered by stock prices, exchange rates, or interest-rate variations can be of the same order of magnitude as current-account surpluses or deficits. In the US case, for instance, external assets mostly consist in foreign-currency-denominated stocks and foreign direct investments, while external liabilities are mostly made up of US-dollar-denominated debt instruments. When the US dollar depreciates in nominal terms, assets are revalued in dollar terms while liabilities are unchanged, and the net foreign-asset position improves. A dollar depreciation thus has a double impact on the net foreign position: First through the current account, then through valuation effects. Between 2001 and 2004, the dollar depreciation, further helped by a better stock price performance abroad than in the US, supported the net foreign position of the US, counteracting the impact of accumulated current-account deficits. As a result, the US net external position could stabilize (figure 5.6) even though the current-account deficit was around 5% of GDP each year. Conversely, the marked appreciation of the dollar in 2008 had a negative impact on the US net foreign asset position, which, concomitant with sky-scraping public deficits, temporarily raised some fear that a dollar crisis could follow the bank crisis.

In order to account for valuation effects, the portfolio-choice model needs amending so as to distinguish gross assets and liabilities. This is done in box 5.7, based on a model developed by Blanchard et al. (2005), who conclude that the dollar depreciation required to reach a given current account target is reduced by a third when valuation effects are accounted for.

In the case of emerging market economies, assets are mostly denominated in domestic currency whereas liabilities are denominated in foreign currencies. Hence, a depreciation of the domestic currency has a negative impact on the net foreign asset position. This makes these countries especially vulnerable to financial crises. During the 2008–09 financial crisis, Central and Eastern European countries (CEECs) temporarily suffered from a *sudden stop**[27] in foreign financing. Since foreign-denominated debts were widespread, especially in Baltic countries, this created a very dangerous situation where the lack of capital inflows would trigger a depreciation of the local currency that, in turn, would make foreign-currency indebted domestic agents insolvent. The International Monetary Fund had to provide emergency financing to several

27. This expression was introduced by Guillermo Calvo to refer to a sudden halt in foreign capital inflows.

of these countries and international institutions persuaded the banks to roll over their credits to the region (on the 2007–09 crisis, see chapter 8).

Box 5.7 The US Current Account and the Dollar: A Model with Valuation Effects

This model is taken from Blanchard et al. (2005). There are two countries: The US and the rest of the world. Capital is perfectly mobile and there is only one type of financial assets (say Treasury bonds). Let S_t be the exchange rate of the dollar at the end of year t (S_t is defined as the foreign currency value of one dollar and it increases when the dollar appreciates) and r and r^* be the respective yields on dollar and foreign-currency-denominated assets, which we suppose constant over time. V_t and V_t^* denote the stock of dollar and foreign-currency-denominated securities at the end of year t, F_t the net foreign debt position of US residents vis-à-vis the rest of the world (in dollars), W_t and W_t^* the wealth of US residents expressed in dollars and of foreigners expressed in foreign currency, B_t and B_t^* the US and rest-of-the-world current accounts of year t respectively. The following accounting relations apply:

$$W_t = V_t - F_t \qquad W_t^* = V_t^* + S_t F_t \qquad B_t^* = -B_t/S \tag{B5.7.1}$$

We first consider the case of a constant portfolio allocation: α is the share of US wealth invested in US assets and α^* the share of foreign wealth invested in foreign assets. We assume $\alpha + \alpha^* > 1$: On average, there is a preference for assets denominated in the home currency of each investor (home bias). From Walras's law, if the dollar market clears, then the foreign currency market clears as well:

$$\begin{aligned} V_t &= \alpha W_t + (1-\alpha^*) W_t && \text{(dollar market equilibrium)} \\ V_t^*/S_t &= (1-\alpha) W_t + \alpha^* W_t^* && \text{(foreign currency market equilibrium)} \end{aligned} \tag{B5.7.2}$$

In particular, the US foreign debt is the difference between foreign investments in the US and US investments abroad: $F_t = (1-\alpha^*) W_t^*/S_t - (1-\alpha) W_t$. From (B5.7.2) and (B5.7.1), one obtains the equilibrium of the dollar market:

$$V_t = \alpha \left(V_t - F_t\right) + \left(1-\alpha^*\right) \left(\frac{V_t^*}{S_t} + F_t\right) \tag{B5.7.3}$$

and finally the exchange rate S:

$$S_t = \frac{(1-\alpha^*)\ V_t^*}{(1-\alpha)\ V_t + (\alpha + \alpha^* - 1)\ F_t} \tag{B5.7.4}$$

An increase in foreign debt F or an increase in the supply of dollar-denominated assets V (say, as a result of fiscal deficits) both depreciate the dollar. The dynamics of external debt is:

$$F_t - F_{t-1} = rF_{t-1} - B_t + (1-\alpha)(1+r)\left(1 - \frac{1+r^*}{1+r}\frac{S_{t-1}}{S_t}\right) \times \left(V_{t-1} - F_{t-1}\right) \quad \text{(B5.7.5)}$$

The variation of foreign debt is the sum of three items: Interest payments to foreigners, "new" debt resulting from the current-account deficit, and a valuation effect depending on the yield differential and on the dollar appreciation and depreciation. In particular, a dollar depreciation ($S_t < S_{t-1}$) reduces foreign indebtedness (since $V_{t-1} - F_{t-1} > 0$).

Let us now drop the simplifying assumption of a constant portfolio structure and move closer to a portfolio-choice model of the type presented in box 5.6. Assume that the proportions α and α^* invested at home depend on the expected relative yield on dollar assets R, and on an exogenous variable x representing the preference for dollar assets over foreign assets (say, because of the liquidity and diversification of the dollar market). We also suppose that the current account depends on the exchange rate and on an exogenous variable z representing the preference for US products over foreign products:

$$\begin{aligned} &\alpha = \alpha(R, x) && \text{with } \partial\alpha/\partial R > 0,\ \partial\alpha/\partial x > 0 \\ &\alpha^* = \alpha^*(R, x) && \text{with } \partial\alpha^*/\partial R < 0,\ \partial\alpha^*/\partial x < 0 \\ &B = B(S, z) && \text{with } \partial B/\partial S < 0,\ \partial B/\partial z > 0 \end{aligned} \quad \text{(B5.7.6)}$$

We can rewrite equations (B5.7.3) and (B5.7.5) into the two following equations: Portfolio balance (PB) and current account balance (CB):

$$V_t = \alpha\left(R_t, x\right)\left(V_t - F_t\right) + \left(1 - \alpha^*\left(R_t, x\right)\right)\left(\frac{V_t^*}{S_t} + F_t\right) \quad \text{(PB)}$$

$$F_t = (1+r)F_{t-1} - B\left(S_t, z_t\right) + \left(1 - \alpha\left(R_t, S_t\right)\right)(1+r) \times \left(1 - \frac{1+r^*}{1+r}\frac{S_{t-1}}{S_t}\right)\left(V_{t-1} - F_{t-1}\right) \quad \text{(CB)}$$

Along a steady-state path where all variables are constant over time, and assuming further that $r = r^*$, both equations produce a decreasing relationship between external debt F and the exchange rate S (see figure B5.7.1). It can be shown that the portfolio balance locus is steeper than the current account balance locus for plausible values of the parameters. The steady-state equilibrium is located at point E. The figure can be used to study the effect of exogenous changes in preferences.

The authors suggest the following explanation for the depreciation of the dollar in the mid 2000s: First, a shift in portfolio preferences toward dollar-denominated assets (a permanent increase in x) shifted the portfolio balance locus to the right, with a new steady state E'; also, a higher preference for foreign goods (a permanent decrease in z) shifted the current account balance locus down, with a new steady state E''. In both cases, the steady-state implication is more external debt and a weaker dollar. The reason is that a higher external debt requires larger interest payments, and thus a larger trade surplus, and thus a weaker currency.

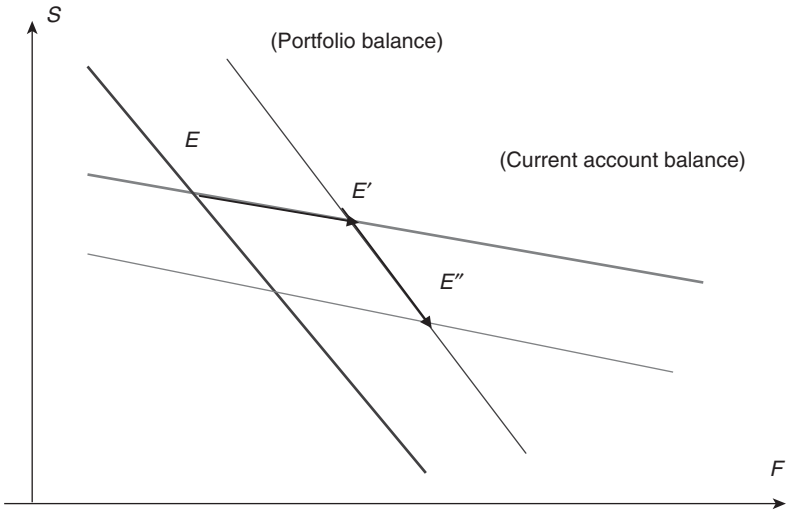


Figure B5.7.1 Exchange-rate adjustment.

For the year 2003, Blanchard et al. find $W = \$35$ trillion, $W^*/S = \$36$ trillion, $F = \$2.7$ trillion, and therefore $V = \$37.5$ trillion and $V^*/S = \$33.3$ trillion. They also find that $\alpha = 0.77$ and $\alpha = 0.70$. Based on these numbers, a 15% depreciation of the dollar improves the US current account balance by 1.4% of GDP, of which one-third is due to valuation effects. They conclude that the required dollar depreciation to reach a given current account target is reduced by a third when valuation effects are accounted for.

5.2.2 Exchange-rate regime choice

As already mentioned, flexible exchange rates are unstable and prone to large and persistent deviations from their fundamental values, introducing uncertainty and distorted relative prices which can impact negatively on GDP growth. However, fixing the exchange rate involves abandoning a

macroeconomic adjustment variable, which may involve more persistent disequilibria or more adjustment needs for other variables. Furthermore, conventionally fixed exchange rates are prone to costly currency crises. Finally, a group of countries facing mostly common shocks may see an advantage in fixing forever their bilateral exchange rates as a way of coordinating their monetary reactions to these shocks. The gain from moving to, say, a monetary union, will be higher the more integrated these countries are, due to the reduction in transaction costs. On the whole, the theoretical arguments that need to be scrutinized when choosing an exchange-rate regime point in various directions, and it is difficult to encompass all of them in a single model. This is why the literature on exchange-rate regime choice remains dominated by simple cost–benefit analyses pioneered by Robert Mundell in the early 1960s.

a) The theory of optimum currency areas

The theory of *optimum currency areas** was introduced in 1961 by Robert Mundell and developed by Ronald MacKinnon (1963) and Peter Kenen (1969). It clarifies the circumstances under which a group of countries should jointly enter a monetary union. The strength of Mundell's argument is that it applies to regions as well as countries, leading to a complete rethinking of monetary geography. In his 1961 paper, Mundell suggested that monetary union would make more sense between the Eastern or Western regions of the US and Canada than within the US and Canada themselves (box 5.8). Mundell's theory also provided the intellectual ground for Europe's economic and monetary union, which was established as a political goal at the Hague Summit of the European Heads of State and Governments in December 1969, and eventually launched 30 years later, in 1999.

Box 5.8 Robert Mundell's Optimal Currency Areas

"A Theory of Optimum Currency Areas" is the title of an article published by Robert Mundell in the *American Economic Review* in 1961. At that time, most currencies had fixed rates under the Bretton Woods arrangement, but the Canadian dollar was floating. A debate was raging between advocates and opponents of floating exchange rates. Mundell, a Canadian economist, cut the Gordian knot by suggesting that this could only be decided on a case-by-case basis. He went further, suggesting that monetary unions would not necessarily coincide with political borders.

Consider the US and Canada, Mundell said. For the sake of simplicity, suppose the Eastern regions of both Canada and the US are specialized in car making, while the Western regions of both countries are specialized in lumber products. Suppose also that the labor force cannot move easily from one coast to another.

Mundell then considers a positive productivity shock in the car industry that creates an excess supply of cars and an excess demand for lumber products (because of the higher demand from car-industry workers). This is a typical asymmetric shock. It causes a dilemma for both the US Federal Reserve and the Bank of Canada. Tightening monetary policy in both regions would curb inflation in the West but increase unemployment in the East. Conversely, cutting interest rates would stimulate employment in the East but lead to overheating in the West. In fact, reorganizing central banks so that there would be an "Eastern dollar" and a "Western dollar" rather than a US dollar and a Canadian dollar would be more suited to this type of economic shock. The central bank of the West could raise interest rates, the central bank of the East could lower them, and the Western dollar would appreciate against the Eastern dollar. Another answer to the dilemma would be labor mobility: Workers would then move from East to West and supply would adjust in both regions (see figure B5.8.1). Absent labor mobility, there is a need for a monetary policy that meets the geography of product specialization.

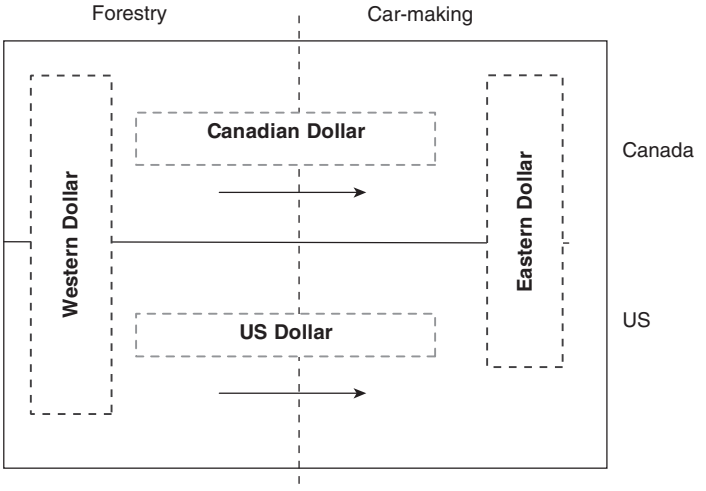


Figure B5.8.1 Monetary borders in North America.

Let us now push the reasoning one step further. Since there are also North–South asymmetrical shocks, creating four currencies and as many central banks could be justified. At the extreme, in order to face all kinds of asymmetric shocks, it would be appropriate that each company and each household issue their own private currency. This would amount to reintroducing barter, and this would not be optimal since it would create a lot of transaction costs. Thus, the borders of the optimum currency areas result from a trade-off between the need to face asymmetric shocks and the need to limit transaction costs.

Mundell's thinking goes as follows. Monetary union brings microeconomic gains because it reduces microeconomic uncertainty and it saves the cost of foreign-exchange transactions. These gains increase with the intensity of trade, i.e., with the level of economic integration. As for the cost of monetary union, it stems from the loss of the exchange rate as an adjustment variable to stabilize aggregate demand. This cost is higher when countries face *asymmetric shocks**, i.e., shocks that affect them in a different way. In this case, an exchange-rate movement instantaneously changes all prices and costs relative to the rest of the world, which helps in adjusting to the shock when prices and costs themselves are sluggish. *Symmetric shocks**, in contrast, affect all countries in the same way and do not require any exchange-rate adjustment. In practice, it is difficult to find a shock that would be strictly symmetric. For the euro area, a fall in US growth may be viewed as mostly symmetric. Some countries—like Ireland—are however more exposed than others—such as Austria—to a downturn in US imports; nevertheless, the US cycle translates into a series of shocks to the euro area that are more symmetric than, say, a fall in world demand for wine, which would mainly affect France, Italy, and Spain. The crucial point for euro area countries is to figure out whether the shocks they face are mostly symmetric or mostly asymmetric.

What if the exchange rate is fixed, Mundell goes on asking, and an asymmetric shock occurs? The answer is price and wage adjustment. Lower prices in a recession-hit region can rebalance demand without the need for an exchange-rate movement. Note that this is no more than a real exchange-rate depreciation. If prices and wages are inflexible, there can always be an adjustment of quantities, in particular through the mobility of labor from one region to another. This is generally the case within the US, where asymmetric shocks do not greatly affect prices and wages but cause workers to move from one state to another (Blanchard and Katz, 1992). It is more problematic within the euro area, due to social, cultural, and linguistic barriers to cross-country migrations, even though other adjustment mechanisms may play a role, such as workers moving in and out of the labor market.

The distinction between symmetric and asymmetric shocks is thus key to understanding monetary unions. However, any judgment based on the nature of shocks *before the monetary union has been put in place* is deemed to be fragile. The nature of shocks can evolve endogenously as a result of the existence of a monetary union. Paul Krugman (1993) first made this point in the case of the US. American states are strongly specialized: The "Corn Belt," the "Rust Belt," or California's Silicon Valley are examples of such specialization. According to Krugman, this has been favored by the US's being a single monetary area: With low barriers to interstate trade, companies could locate freely to exploit agglomeration spillovers such as a specialized labor force, infrastructures, or sub-contractor networks.[28] Being highly specialized, states are more prone to asymmetric shocks: A fall in car prices hits Michigan

28. The geographic distribution of industries is further discussed in chapter 6.

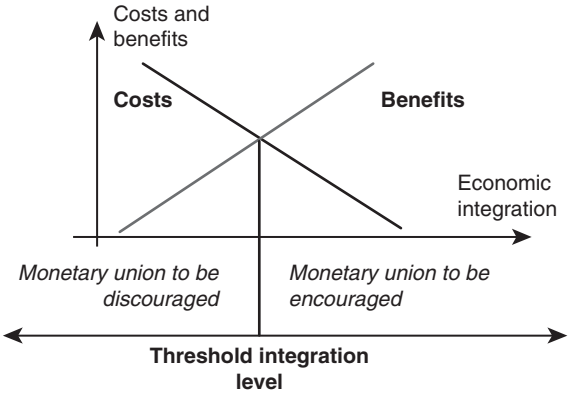


Figure 5.14 Costs and benefits of a monetary union.

more than any other region. European monetary union, Krugman concluded, would increase rather than decrease economic heterogeneity in Europe.

This vision does not meet unanimity. On the one hand, Frankel and Rose (1998) and Rose (2000) found that monetary unions have led to more trade integration, which implies that the benefits from monetary union are also partly endogenous.[29] On the other hand, Fontagné et al. (2006) showed that European integration has boosted intra-industry trade rather than inter-industry trade, reducing the likelihood of asymmetric shocks. Indeed, monetary union would preserve or even raise product diversification in member countries, which has been shown by Kenen (1969) to be a major criterion of optimal currency areas.

Mundell's cost–benefit approach is illustrated in figure 5.14, inspired by Paul Krugman. The benefits of monetary union are an increasing function of integration. The costs can be an increasing or decreasing function of integration, depending on whether integration raises or reduces the asymmetry of shocks. The threshold level depends on the positions of the "gain" and "cost" locus. More flexible prices and wages reduce the usefulness of the exchange rate as an adjustment instrument: The "cost" locus moves downward, as does the threshold level of integration. At the limit, in an economy with perfectly flexible prices and wages, the exchange rate does not matter at all. In contrast, in countries that are prone to asymmetric shocks (such as economies in transition to market economy), the "cost" locus is higher, as is the threshold level of integration. For these countries, the exchange rate remains an important economic policy instrument.

29. Andrew Rose triggered an intense controversy by assessing, based on past monetary union experiences, that, *ceteris paribus*, such a union would lead to a threefold increase in trade between members of the union. Later estimates provided lower figures, but still a significant increase in trade following monetary union. See Baldwin (2006).

b) Monetary or real shocks?

In addition to the distinction between symmetric and asymmetric shocks, the nature of the shocks is important in assessing the usefulness of flexible exchange rates. In 1970, William Poole devised a simple, Keynesian framework to compare the effectiveness of money supply and the interest rate as stabilization instruments for aggregate demand, depending on whether the economy faces nominal or real shocks. *Nominal shocks** are shocks to the money supply or to the velocity of money (see chapter 4) while *real shocks** are shocks to consumption or exports which directly affect aggregate demand. Albeit devised in a closed economy setup, the Poole model can easily be extended to study the choice of an exchange-rate regime. It can then be shown (see box 5.9) that, in an economy dominated by monetary shocks, a fixed exchange rate allows for greater output stability. This is the case, for example, if private agents hold liquidities in foreign currency, or if the financial system is unstable. In an economy dominated by real shocks, as is the case in advanced industrial economies, a flexible exchange rate stabilizes output better.

Box 5.9 Nominal versus Real Shocks: Extending Poole's Model

We use the small, open economy IS–LM model, also known as the Mundell–Fleming model, presented in chapter 3. We assume that capital is perfectly mobile and, for the sake of simplicity, that expectations are static. The domestic nominal interest rate is thus always equal to the world interest rate.

A positive shock to money supply moves the LM curve *ex ante* to the right (left-hand-side panel of figure B5.9.1): Output increases from its initial value Y_0 and the interest rate declines. In a fixed exchange-rate regime, the central bank has to sell foreign currency to clear the foreign-exchange market. By so doing, it withdraws domestic currency from circulation and brings domestic money supply down to its initial level. The LM curve goes back to its initial position and output has eventually not increased: $Y_{fix} = Y_0$. In a flexible exchange-rate regime, in contrast, the foreign-exchange market is cleared by nominal exchange-rate depreciation. Depreciation in turn stimulates net exports, pushing the IS curve to the right: The initial impact of the shock is thus magnified by the exchange-rate movement: $Y_{flex} > Y_0$.

In the case of a positive real shock, like a fall in the saving rate, the effect is quite the opposite (right-hand-side panel of figure B5.9.1). The shock moves *ex ante* the IS curve to the right, pushing both output and the interest rate up. In a flexible exchange-rate regime, there is no *ex post* impact on output because the exchange rate appreciates and weighs on

foreign trade: $Y_{flex} = Y_0$. In a fixed exchange-rate regime, the impact of the shock is magnified by the central bank, which buys foreign reserves to counter the appreciation of the currency and thus increases the quantity of money in circulation, which leads to a further output expansion: $Y_{fix} > Y_0$.

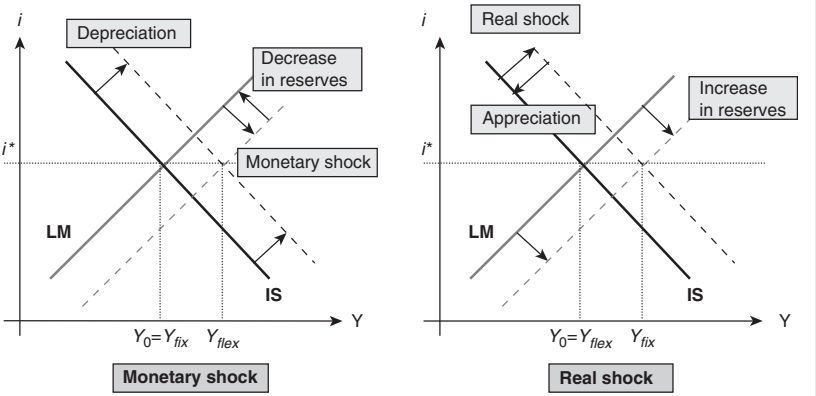


Figure B5.9.1 Exchange-rate regime choice in the Poole model.
Reading:

- Left-hand side: A positive monetary shock translates the LM curve to the right. In a flexible exchange rate regime and if the Marshall–Lerner condition holds, income increases from Y_0 to Y_{flex} because the exchange rate depreciates, moving IS to the right. In a fixed exchange-rate regime, income ultimately remains in Y_0 because LM moves back up as a result of lower foreign-exchange reserves to finance the current-account deficit.
- Right-hand side: A positive real shock translates the IS curve to the right. In a fixed exchange rate regime, income increases from Y_0 to Y_{fix} because official reserves have increased to prevent the appreciation of the currency. In a flexible exchange rate regime, income ultimately remains in Y_0 because the IS curve moves back to the left as a result of the appreciation of the exchange rate.

This result has important implications for emerging economies, and especially transition economies: Before the credibility of the domestic currency is well-established and the financial system has developed, nominal shocks are likely to be an important phenomenon and a fixed exchange rate can help stabilize demand. In a later stage, nominal shocks become less important and the country should move to a more flexible regime. In practice, most transition economies started with fixed exchange-rate regimes at the beginning of the transition before moving to floating rates. Interestingly, only the smallest—hence more open—new EU member countries such as Slovenia, Slovakia and the Baltic countries have made significant steps toward monetary union so far.

c) Risk sharing

There are ways to react to shocks other than price and wage adjustment. In a second paper published in 1973, Robert Mundell argues that a monetary union provides automatic insurance against asymmetric shocks because economic agents share a common pool of money. More generally, international financial integration provides a device to smooth consumption in each country or region, due to portfolio diversification: If the domestic economy is hit by, say, a negative, specific productivity shock, then domestic income falls, but households will continue to earn returns from their financial holdings invested abroad. Macroeconomic risk sharing can also be achieved by means of *fiscal federalism**, that is, compulsory transfer of fiscal resources across geographical constituencies, such as regions or countries. If the domestic economy is hit by a negative shock, then its contribution to the federal budget is reduced and it receives transfers from it. On the whole, financial integration as well as the existence of a federal budget are important criteria for forming a currency area. By decomposing the disposable income of each US state in order to highlight the proportion of its net income originating from other states (dividends, interests, rental income payments across state borders, federal transfers, and taxes), Asdrubali et al. (1996) find that 39% of shocks to gross state product are smoothed by portfolio diversification, 13% by the federal budget, and 23% by credit markets, over the 1963–90 period (see box 5.10). Hence, only 25% of idiosyncratic shocks remain unsmoothed.

In contrast, Sørensen and Yosha (1998) find no evidence of risk sharing through portfolio diversification across OECD countries nor across EU countries over the same period. High portfolio diversification within countries but not across countries is consistent with the Feldstein–Horioka puzzle. Due to its small size (1.27% of the region's GNP) as well as its focus on agriculture and structural expenditures, the EU budget does not play any role in risk sharing across member countries. It can be concluded that the cost of idiosyncratic shocks to EU member countries is high due to the lack of any EU-wide risk-sharing scheme. However, free capital movements were not legally enforced in the EU until 1990, and in the early 1990s full financial market integration was far from being achieved. Updated results reported by Kalemli-Ozcan et al. (2004) find that 6% of EU country-specific shocks were smoothed through the international capital-income-flows channel over the 1993–2000 period. The importance of capital market integration for the sustainability of monetary integration has also been recognized by African countries sharing the CFA franc (see box 5.14 below) and by the ASEAN-plus-three grouping in East Asia,[30] with the so-called "Asian bond-markets" and "Chiang-Mai" initiatives of 2003 toward monetary integration in this region (Henning, 2009).

30. The ASEAN-plus-three group brings the ten ASEAN countries (Brunei, Burma, Cambodia, Indonesia, Laos, Malaysia, Philippines, Singapore, Thailand, and Vietnam) together with China, Korea, and Japan. It has become the main cooperation forum in East Asia.

Box 5.10 Measuring the Extent of Risk Sharing

Asdrubali et al. (1996) have measured the extent of interstate risk sharing in the US. Their measure is based on the following identity:

$$C_i = \frac{C_i}{D_i} \times \frac{D_i}{I_i} \times \frac{I_i}{Y_i} \times Y_i \qquad \text{(B5.10.1)}$$

where C_i denotes consumption in state i, D_i disposable income, I_i income and Y_i GDP, all on a per-capita basis. Equation (B5.10.1) states that consumption per capita may be unrelated to GDP per capita thanks to the use of credit (first term), to the federal tax and benefit system (second term), and to capital income (third term). Asdrubali et al. regress the log-variation of each of these three terms on the log-variation of GDP per capita. For the third term, for instance:

$$\Delta \log Y_{it} - \Delta \log I_{it} = \nu_t + \beta \Delta \log Y_{it} + u_{it} \qquad \text{(B5.10.2)}$$

where ν_t denotes time-fixed effects and u_{it} is the residual. The coefficient β can be interpreted as the percentage of state i's GDP per capita variations that are *not* passed on to state i's income per capita, thanks to the counteracting impact of capital income. Based on real per-capita gross state product in 50 US states over 1963–90, Asdrubali et al. find β to be as high as 0.39, i.e., a 1% drop in GDP per capita leads to a fall in income per capita of only $(1 - 0.39) = 0.61\%$. Similarly, they estimate a β coefficient of 0.13 for smoothing through the federal budget and 0.23 for smoothing through the credit market, which can either come from international lending or from a purely domestic credit market.

Using a similar methodology, Sørensen and Yosha (1998) find international risk sharing through capital markets to be insignificant both across OECD and across EU countries. Using a slightly different methodology, Mélitz and Zumer (1999) find a much smaller amount of capital risk-sharing across EU countries.

5.2.3 Models of currency crises

Up to this point, we have focused on steady-state exchange-rate regime choice, leaving aside the issue of regime change. But we have seen that what is appropriate for a country at a given point in time may not remain so forever. Governments may have to change exchange-rate regimes; if they fail to do so, market forces may produce a *currency crisis**, forcing them to modify their exchange-rate regime. Economists are often derided for their facility in providing sophisticated explanations for the latest currency crises while not being able to predict the next one. This occurs because the nature of crises changes over time. Until the 1980s, capital did not move freely. Sustainability of a fixed exchange rate would depend on the availability

of foreign-exchange reserves. The driving force of currency crises was the current account: Crises would typically erupt in countries with high aggregate-demand growth, leading to unsustainable current-account positions. In the 1990s and 2000s, with free capital movement and the deepening of global financial markets, currency crises increasingly originated in the financial account because of excessive short-term foreign currency debt.

a) First-generation currency-crisis models

The inconsistency between domestic economic policy and the exchange-rate regime is at the core of *first-generation currency-crisis models** (Flood and Garber, 1984; Krugman, 1979).[31] Imagine a country with a fixed exchange rate against the US dollar where monetary aggregates grow faster than in the US. As a consequence nominal demand grows faster and the country experiences a current-account deficit. In order to maintain its exchange rate, it has to draw on its foreign-exchange reserves. The insight of first-generation models is to show that the crisis will typically occur *before* reserves are exhausted.

The mechanism is comparable to a run on bank deposits (see chapter 4, box 4.17). In a bank run, depositors rush to the counter to withdraw their deposits because they anticipate that the bank will not be able to refund all of them. Even if they have no prior view on the bank insolvency, they just *have to* rush because they do not want to be last at the counter. In a currency crisis, it is just about the same: In a speculative attack, financial market participants rush to change their domestic currency holdings into foreign currency because they fear that the central bank may not have enough reserves to convert the whole domestic currency outstanding at the initial fixed rate. The only way to make sure the run cannot happen is for the central bank to hold in its vault as much foreign currency as the total domestic currency in circulation: This is the basis of the currency-board system, presented earlier in this chapter.

The insight of the Krugman (1979) model is that the crisis occurs at a precisely determined moment, namely when the "shadow" floating exchange rate, i.e., the exchange rate that would prevail in a floating-exchange system, becomes lower than the actual fixed rate. Before this point has been reached, there is no expected gain in selling the currency. After it has been reached, profit is certain and all market participants sell.

First-generation models thus consider the crisis as the rational reaction of market participants to an inappropriate exchange-rate regime. A good example is Pakistan in 2008 (figure 5.15). Monetary growth was fueled by unbridled public spending and direct government borrowing from the central bank and was much higher in Pakistan than in the US. The only way to sustain the fixed US dollar–Pakistani rupiah exchange rate was for the State Bank of Pakistan to sell US dollars and buy rupiahs. As a consequence, foreign-exchange reserves were gradually depleted and, in the summer of 2008,

31. The model draws on an article by Salant and Henderson (1978) on the exhaustion of gold reserves.

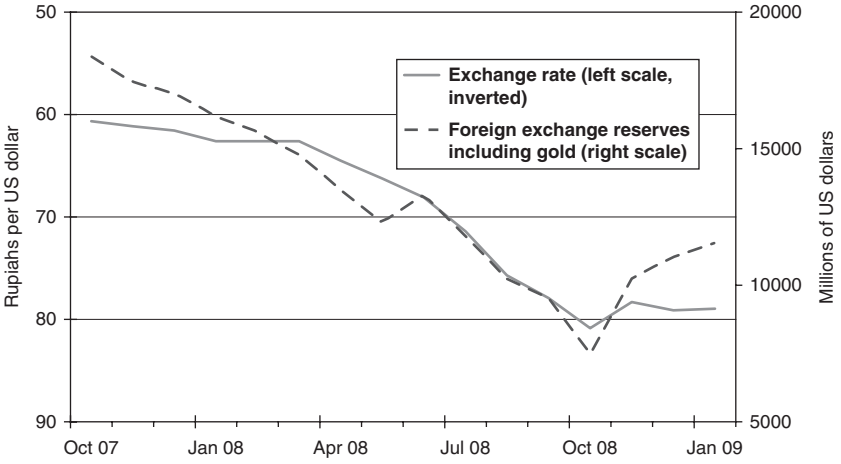


Figure 5.15 A balance-of-payment crisis: Pakistan, 2008.
Source: State Bank of Pakistan.

market participants triggered the crisis by selling in a rush their remaining rupiahs. The exchange rate stabilized only when the IMF and the Asian Development Bank extended foreign-currency-denominated loans to the government, and reserves could then be gradually replenished.

b) Second-generation currency-crisis models

The mechanism described in the first-generation models is well fitted to countries with high inflation or an unsustainable current-account deficit. However, consider the European exchange-rate crises of 1992–93. At that time, European countries were undertaking restrictive monetary policies in the wake of German unification, which had caused an expansion of aggregate demand; some of them registered current-account *surpluses* and none of them was lacking foreign-exchange reserves. First-generation currency-crisis models fall short of explaining what happened in Europe. What is needed here is not a monetary model, but a model with short-term nominal rigidities, linking the exchange rate and the real economy. What is also needed is an explanation of why a crisis can occur even in the absence of unsustainable trends. The contribution of *second-generation currency-crisis models** was twofold: first, to show that devaluation expectations can be self-fulfilling: The mere expectation that the government will devalue can force it to do so; second, to show that crises may result from the interaction between rational market expectations and the optimizing behavior of governments.

There may be several, unrelated reasons why financial market expectations can feed back onto the real economy. Inflationary expectations provide a first explanation (Jeanne, 1996): If private agents expect a devaluation, they will ask for higher wages to compensate for their expected loss of purchasing power. In a supply-driven economy, this pushes unemployment up unless

the government actually devalues to restore corporate profits. This is the open-economy version of the Barro–Gordon time-consistency dilemma (see chapter 4, box 4.9). Box 5.11 presents a simplified model along these lines. Another explanation, probably more relevant to the European experience, is demand-driven (Obstfeld, 1994). Under uncovered-interest-rate parity, devaluation expectations push nominal interest rates up. This penalizes aggregate demand and induces the government to devalue to revert to the initial output level. In short, expectations of a crisis may be enough to cause it: Such crises are *self-fulfilling currency crises**.

Box 5.11 A Self-Fulfilling Currency Crisis

This model was proposed by Jeanne (1996) as an attempt to explain the European currency crises of the early 1990s. We consider a small economy. Domestic prices are determined by purchasing power parity. The exchange rate is fixed but can be devalued by a fixed percentage δ. A surprise devaluation would cause a short-term decrease in unemployment through lower real wages (remember the Barro–Gordon model of box 4.10). However, it would impair the reputation of the central bank. The government thus faces a dilemma.

The central bank minimizes the quadratic loss function: $L = u^2 + cz$, where u designates the unemployment rate, c the (fixed) political cost of a devaluation and z is equal to 1 in the event of a devaluation and 0 otherwise.

The unemployment rate is sluggish and is negatively affected by unexpected inflation, i.e., the difference between actual inflation π and expected inflation π^e:

$$u = \rho u_{t-1} - \lambda(\pi - \pi^e) \text{ with } 0 < \rho < 1, 0 < \lambda < 1 \qquad \text{(B5.11.1)}$$

Under purchasing power parity, expected inflation is equal to the expected rate of exchange-rate devaluation. Expected inflation rate is thus either 0 (no devaluation) or δ (devaluation). In this setting optimal monetary policy depends on expectations:

- If private agents do not expect any devaluation, then expected inflation is zero and realized inflation is either 0 or δ depending on whether or not there has been a devaluation. The loss function is either $L_0^0 = \left(\rho u_{t-1}\right)^2$ (no devaluation) or $L_0^d = \left(\rho u_{t-1} - \lambda\delta\right)^2 + c$ (devaluation). The government rationally chooses to devalue if $L_0^d < L_0^0$, i.e., if:

$$\Phi < -\lambda\delta \quad \text{with } \Phi = \frac{c}{\lambda d} - 2\rho u_{t-1} \qquad \text{(B5.11.2)}$$

  This condition is more likely to be met if the cost of devaluing c is low and unemployment inherited from the previous period, ρu_{t-1} is high. Φ summarizes the "fundamentals" of the economy.

- If private agents expect a devaluation, then expected inflation is δ. The loss function is either $L_d^0 = \left(\rho u_{-1} + \lambda\delta\right)^2$ (no devaluation) or $L_d^d = \left(\rho u_{-1}\right)^2 + c$ (devaluation). The government chooses rationally to devalue if $L_d^d < L_d^0$, i.e., if:

$$\Phi < \lambda d \qquad \text{(B5.11.3)}$$

This condition is more easily met than when no devaluation is expected. Figure B5.11.1 summarizes the results. If fundamentals are excellent ($\Phi > \lambda\delta$), there is no devaluation even if agents expect it to happen. If fundamentals are poor ($\Phi < -\lambda\delta$), there is a devaluation even if agents do not expect it to happen. For medium fundamentals ($-\lambda\delta < \Phi < \lambda\delta$), there are two possible equilibria and expectations are self-fulfilling: The economy can shift from one to another (e.g., the government can be forced to devalue) if there is a shift in private expectations.

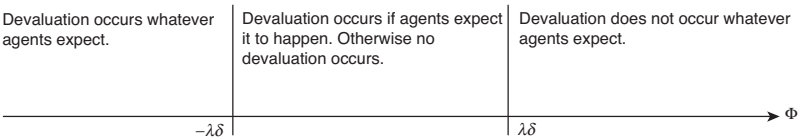


Figure B5.11.1 Fundamentals and expectations in a second-generation currency crisis.

According to second-generation models, countries vulnerable to a currency crisis can be identified by the conjunction of a relatively weak economy and adverse market expectations. The best example is the crisis of the pound sterling in 1992. The pound sterling had joined the European Exchange Rate Mechanism a year earlier, at a central rate against the deutschemark that many observers deemed too high. In 1992, the economy was weak and speculative attacks began to materialize. The Bank of England then raised its interest rate, but only moderately, sending the signal that the government was not ready to worsen the recession to defend the currency. After massive foreign-reserves losses, participation in the Exchange Rate Mechanism was suspended and the pound began to float. The Chancellor of the Exchequer, Norman Lamont, subsequently declared that he had been "singing in his bath" that evening. Change of the exchange-rate regime, prompted by the currency crisis, had lifted a burden from the government's shoulders.

c) Third-generation currency-crisis models

Starting from July 1997, Thailand, then other East Asian economies, Brazil and Russia had to give up their fixed exchange rate against the US dollar. Once again, the prevailing currency-crisis model had not helped anticipate crisis.

The main channel of transmission of an exchange-rate devaluation to the real economy was now through the financial sector (Corsetti, 1998). All these countries had in common US-dollar-denominated and/or short-maturity external debt (as an example, foreign banks had granted loans to their Thai subsidiaries to invest in the then-booming domestic-housing sector). Any devaluation of the currency would increase the burden of foreign currency-denominated debt and/or cause a "sudden stop" of capital inflows. A similar mechanism was at play in 2008 in several non-euro-area European economies such as Hungary, Iceland, and Latvia (which could eventually maintain its fixed exchange rate), which were heavily dependent on short-term, euro-denominated external financing.

A banking crisis can cause a currency crisis: Market participants anticipate that the central bank will lend fresh money to distressed banks, in its role as lender of last resort (chapter 4). The expected liquidity injection generates by itself a devaluation expectation. A currency crisis can also cause a banking crisis, when banks' debts are denominated in foreign currency. A vulnerable banking sector (due to a high proportion of nonperforming loans, for example) reinforces risks in both directions. *Third-generation currency-crisis models** (Krugman, 1999) have focused on the "twin" banking and currency crises.

Third generation models have also sought to explain *contagion effects** at work in the late 1990s and again in the late 2000s, particularly at the regional level, in Europe and in Asia. Contagion in East Asia is apparent in figure 5.16. In 1997, Thailand broke its quasi-fixed link to the dollar and was followed closely by Indonesia and Malaysia, then Korea. Masson (1999) has proposed an explanation based on regional trade. He uses a first-generation currency-crisis model, adding a regional dimension: The trade performance of a country depends on the exchange rate of its regional competitors. If one of them devalues, its trade balance will deteriorate and it will have to intervene in the market to defend its fixed parity. Knowing this regional dependency, market participants may find it optimal to attack the currency in a preemptive way. More generally, contagion can result from a change in investor sentiment following an unexpected event in a given country. In July 1997, the Thai baht crisis was a wake-up call to investors who had forgotten that investment in emerging market economies was inherently risky. The upward revision in the price of risk was immediately built into the risk premia paid by *all* emerging market economies, not only Thailand (remember the risk premium in the portfolio model of box 5.6). Finally, an additional channel of contagion stems from the international investors' budget constraint: In order to offset losses in a given country, they sell assets that have not depreciated yet, throwing other countries into the crisis.

Currency-crisis models can be tested by reality, as should all theoretical models. Probabilistic methods or econometric estimates are used to assess the predictive power of the theoretical model (box 5.12). It turns out that the most important variable is the degree of overvaluation of the real exchange

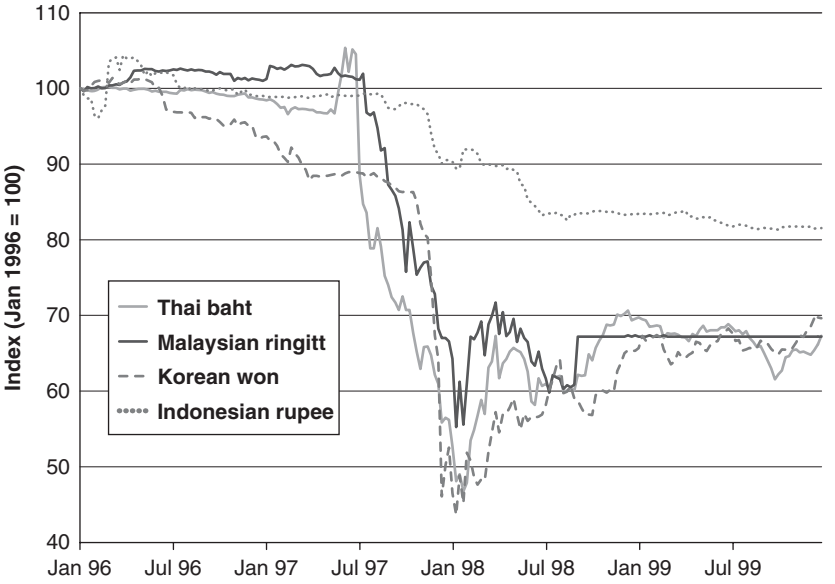


Figure 5.16 Exchange-rate contagion in Asia, 1996–99.
Source: Reuters.

Note: US dollar end-of-week exchange rate, rebased at 100 on 1 January 1996.

rate—which does not come as a surprise and takes us back to the need to produce reliable models of equilibrium exchange rate. The ability of these empirical models to forecast crises is disappointing, albeit it is somewhat improved when regional contagion effects are introduced (Kaminsky and Reinhart, 2003).

Box 5.12 Predicting Exchange Crises

The literature on forecasting exchange rate crises has two main branches.

In the *probabilistic approach* (Kaminsky et al., 1998), the authors select a number of variables likely to be associated with a currency crisis. For each of these variables, they set a threshold value that minimizes the ratio between the number of erroneous signals (i.e., warnings not followed by crises) and the number of correct signals (warnings followed by crises). All variables are then aggregated to build a so-called "crisis probability," with a threshold value triggering a warning signal.

The *econometric approach* aims at modeling explicitly the probability of a crisis.

In so-called "probit" models (Frankel and Rose, 1996), a binary endogenous variable p expresses the probability of a crisis as a function of a set of exogenous variables X: $p = \Phi^{-1}(Z'X)$ where Φ is the standard

normal cumulative function and X is the vector of the explanatory variables. The coefficients Z are estimated from time series. The function is used out of sample to predict the probability of a crisis, with a "warning signal" when p is above a given threshold value.

In traditional econometric models (Sachs et al., 1996), the endogenous variable is continuous. It can be, for instance, a combination of exchange-rate and foreign-reserves variation. The model does not directly yield a crisis probability.

In all cases, explanatory variables are deduced from the three generations of currency crises models. The variables that best predict crises are the following:

- Money aggregates: Growth rate of foreign reserves, level and growth rates of the ratio of M2 to foreign-exchange reserves, growth rate of credit.
- Risks to foreign-exchange reserves: Current account to GDP ratio, real exchange-rate overvaluation, growth rate of exports.
- Fiscal expansion: Fiscal deficit/GDP ratio.
- Cost of maintaining a fixed exchange rate: World interest rate, real exchange-rate overvaluation.
- Risk to the financial account: External debt, share of concessional debt in total debt, share of short-term liabilities, share of debts in foreign currencies.
- Risk of a banking crisis: Nonperforming loans.
- Contagion risk: Crisis probability in neighboring countries.

A major difficulty is that empirical models usually succeed in predicting past crises based on the value of exogenous variables as realized *ex post*, but not based on information available at the time of the forecast (Berg and Patillo, 1999).

5.3 Policies

The first policy choices when it comes to exchange rates are the exchange-rate regime and the degree of capital mobility. In a floating-rate system, policymakers need to decide whether to try to influence the exchange rate or to let it float freely. We successively address these issues. We then reflect on how national policy choices interact with the international monetary system at a regional and at a global level.

5.3.1 Capital mobility and the choice of an exchange-rate regime

In chapter 3, using the Mundell–Fleming model, we established that the preferred output-stabilization policy should depend both on the degree of

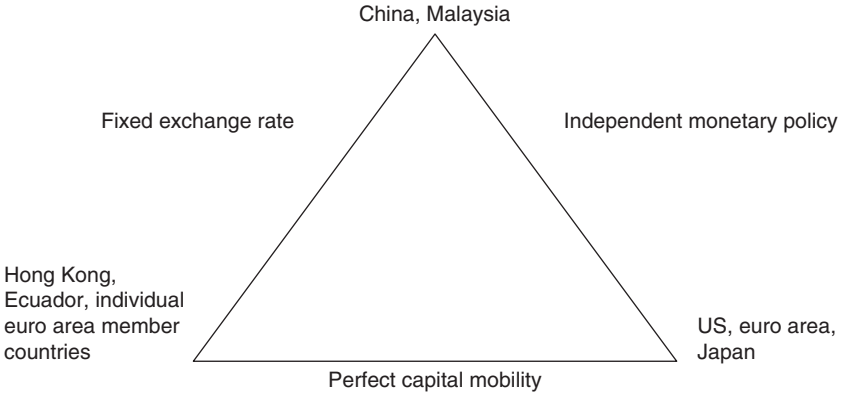


Figure 5.17 Mundell's Impossible Trinity.

capital mobility and on the exchange-rate regime. When capital is mobile and the exchange rate is floating, monetary policy should be preferred to fiscal policy: The impact of monetary expansion on output is magnified by nominal exchange-rate variation triggered by changing interest rates. This is the external transmission channel of monetary policy. In contrast, a fiscal expansion is self-defeating since it pushes interest rates higher and appreciates the exchange rate at the expense of net exports. With mobile capital and a fixed exchange-rate regime, it is quite the opposite: A fiscal expansion is magnified by capital inflows that add to foreign-exchange reserves, generating fresh money creation; a monetary expansion is self-defeating since it lowers interest rates, leading to foreign reserve depletion that withdraws money from circulation.

This has important consequences for the choice of an exchange-rate regime. Remember the Tinbergen rule in chapter 2: There is no need to use two policy instruments to target one objective. When capital is mobile and the exchange rate is floating, the preferred instrument for output stabilization should be monetary policy. Indeed, recovering monetary independence was one major debate in the 1960s, at the time Robert Mundell was writing his optimal currency area theory. Conversely, a fixed exchange rate with free capital movements requires the central bank to dedicate monetary policy to exchange-rate stabilization. Output stabilization then goes through fiscal policy, which is a less reactive tool than monetary policy and involves public debt accumulation.

One popular way to summarize these findings is the *impossible trinity** or *Mundell's triangle**, which states that a country cannot simultaneously enjoy an independent monetary policy, a stable nominal exchange rate, and a perfectly mobile capital (Figure 5.17).

Policymakers need either to choose one summit of the triangle, or find an appropriate trade-off between, e.g., some monetary independence and some exchange-rate stability. The triangle sheds a useful light on the many

contradictions of twentieth-century exchange-rate policies. In the 1930s, European countries faced a contradiction between the Gold Standard fixed exchange rates, capital mobility, and the economic consequences of the war, which called for different monetary policies in different countries. Eventually, they had to adjust their exchange rates or leave the Gold Standard. In the late 1990s, East Asian and Latin American countries faced the same kind of contradiction. Eventually, most of them chose to let their exchange rate float. However, some of them preferred to move to other summits of the triangle. Malaysia retained its fixed rate against the dollar and installed capital controls, while Ecuador gave up monetary independence and dollarized its economy. China, which had not yet liberalized its financial account, did not have to change its policies significantly and resisted the crisis. In the financial crisis of the late 2000s, once again, countries like Latvia experienced the difficulty of maintaining a fixed exchange rate vis-à-vis the euro when ailing Western European banks could withdraw their financial support overnight.

In the remainder of this section, we discuss the criteria that policymakers can use when deciding between the three vertices of Mundell's triangle. These are: The desirability of capital mobility; macroeconomic and microeconomic criteria based on the theory of optimal currency area; credibility gains that can be brought by a fixed exchange rate; and criteria related to international coordination.

a) The pros and cons of capital openness

As can be seen in figure 5.1, there has been a continuous trend since World War II toward free movement of capital. Since 1944, liberalization of capital movements has been encouraged by the IMF to allow for an optimal allocation of world savings and to let money flow from rich capital-abundant countries to poor capital-scarce countries. Important milestones were the demise of the Bretton Woods system in the early 1970s, and then the European Single Act of 1992, which removed all capital restrictions inside the European Union and encouraged free movement of capital with the rest of the world. Throughout the 1990s, capital movement liberalization was part of the "Washington consensus" (see chapter 6) and was imposed by the G7, OECD, and IMF on many emerging nations.[32] This resulted in an unprecedented development of international capital flows, as illustrated in figure 5.18.

The currency crises of the late 1990s in emerging markets dealt a blow to the consensus on free capital mobility. As described in section 5.2, capital liberalization in Asia had unleashed short-term, dollar-denominated capital inflows which had fuelled asset prices and bank lending and had led ultimately to "twin" banking and currency crises. Furthermore, the

32. Financial-account convertibility was famously set as a precondition to the Republic of Korea's OECD accession in 1996. One year later, the economy collapsed in a financial crisis.

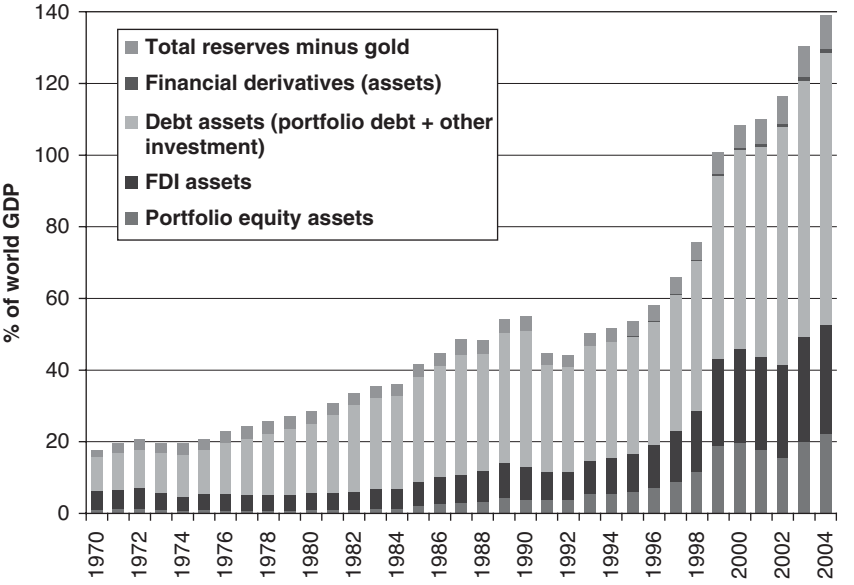


Figure 5.18 Gross foreign assets as a percentage of GDP, world aggregate, 1970–2004. Source: Lane and Milesi-Ferretti (2006).

contribution of capital mobility to long-term economic development was increasingly put into question. In a 1990 paper, Robert Lucas noted that most capital flows were "North–North" flows between rich countries rather than "North–South" flows from rich to poor countries, as standard growth theory would have predicted. One explanation for this paradox is information asymmetry and weak institutions in low-income countries, which prevent capital from being invested there. Modern capital flows are "diversification finance" rather than "development finance" (Obstfeld and Taylor, 2004). Moreover, the welfare gains of financial openness have been found to be much lower than previously thought. Using a neo-classical growth model of the type presented in chapter 6, Gourinchas and Jeanne (2006) found that for non-OECD countries, the welfare effect from switching from financial autarky to perfect capital mobility is equivalent to a permanent increase in consumption of about 1%. This is not much. It is much lower than the impact of the take-off in domestic productivity which has taken place in some of these countries, and it is an order of magnitude lower than the short-term impact of currency crises on output (see section 5.1). The direct impact of financial openness, through lower cost of capital, is much less than its indirect effects through productivity spillovers, external pressure to remove distortions on domestic markets, and better economic institutions. The relationship between openness and long-term growth will be further discussed in chapter 6.

After the Mexican, and then the Asian, crises of the 1990s, it did not come as a surprise that emerging countries reconsidered the merits of capital openness. China, which had continuously resisted opening its financial account, was praised for supporting financial stability in the region. After a failed attempt to implement an orthodox, IMF-led interest-rate increase and budgetary retrenchment, Malaysia fixed the exchange rate at a *higher* level (at 3.80 Malaysian ringgit for a US dollar, the ringgit was appreciated by 10%), cut interest rates, and reinstalled capital controls.[33] With hindsight, judgment on the Malaysian experience is at best mixed (see Kaplan and Rodrik, 2001, for a more-lenient vision). The IMF nevertheless took a more benign approach to capital controls, and its chief economist, Michael Mussa, recognized that:

> The experience in recent financial crises could cause reasonable people to question whether liberal policies toward international capital flows are wise for all countries in all circumstances. The answer, I believe, is probably not. . . . high openness to international capital flows, especially short-term credit flows, can be dangerous for countries with weak or inconsistent macro-economic policies or inadequately capitalized and regulated financial systems.
>
> Michael Mussa (2000), pp. 43–44

Several types of capital controls exist. A first option is administrative control of foreign-exchange operations, with restrictions on the type of operations that can be undertaken. In 2006, 165 countries were complying with Article VIII (sections 2, 3 and 4) of the IMF by-laws, which aim at removing controls on current-account transactions. Most remaining restrictions related to financial transactions, and to payments and transfers on invisible transactions (e.g., remittances), not imports of goods. Nearly all countries were maintaining some controls on capital transactions: Generally, controls on transactions by banks and on financial instruments (sometimes combined with prudential supervision) and on foreign direct investments for national security reasons.

Other options are based on incentives rather than control. The aim is to "throw sand in the wheels" of financial globalization by discouraging short-term capital flows, which are deemed to generate exchange-rate volatility. In the 1990s, the Chilean government forced foreign investors to leave a minimum reserve (*encaje**) at the central bank as an interest-free deposit. In the same spirit, James Tobin proposed a tax on foreign-exchange transactions. Proponents of the *Tobin tax** usually advocate a 0.05% levy on transactions unrelated to a current-account transaction or foreign direct investment.

33. These were: Compulsory transaction through authorized intermediaries to sell ringgit-denominated assets, a 12-month ban on repatriation of ringgit proceeds obtained by foreigners, forced repatriation of ringgits held abroad, etc.

The Tobin tax follows from a remark by John Maynard Keynes on the impact of taxes on stock market liquidity:

> It is usually agreed that casinos should, in the public interest, be inaccessible and expensive. And perhaps the same is true of stock exchanges. . . . The introduction of a substantial Government transfer tax on all transactions might prove the most serviceable reform available, with a view to mitigating the predominance of speculation over enterprise in the United States.
>
> John Maynard Keynes (1936), pp. 159–60

Building on Asian experience, the IMF has moved to advising a sequencing of financial account liberalization in emerging countries, where full liberalization should only come after a sound financial system has developed domestically with well-managed and supervised banks and a low level of nonperforming loans. This is indeed the strategy followed by China.

However, maintaining foreign-exchange controls is not advisable in the long run, for at least two reasons. First, Chilean-type reserves or a Tobin tax are not enough to discourage capital outflows when the expected depreciation of the domestic currency becomes high (see Eichengreen et al., 1995, for a discussion). Second, and more importantly, the volume of foreign-exchange transactions does not stem from speculation but from exchange rate and liquidity risk being processed and redistributed among financial intermediaries and to end-investors. Third, taxes on specific transactions as well as administrative controls can be circumvented through financial innovation or off-shoring to countries which do not impose them. And it is impossible to identify "speculative" foreign exchange transactions that aim at a short-term profit. The same limitations apply to attempts by governments to segment capital markets by imposing capital controls (Garber and Taylor, 1995). Fourth, direct taxes on banks or financial activities are a better way to discourage financial development if deemed excessive. Finally, the only convincing justification for a Tobin tax is purely fiscal: If accepted by all countries, it would be a means to tap a very large tax base and alleviate the tax burden on immobile production factors such as low-skilled labor (see chapter 7).

b) Criteria for exchange-rate-regime choice

As explained in section 5.2, the rationale for joining a monetary union or fixing the exchange rate of one currency against another one results from a trade-off between microeconomic benefits and macroeconomic costs. Microeconomic gains relate to the cost of exchange-rate volatility and to exchange-rate conversion costs. They are particularly difficult to quantify. In the European case, the Emerson report (1992) by the European Commission valued them at between a quarter and half a percentage point of GDP, but this estimation has been questioned. They can be substantially higher for

small, very open economies, which do not trade in their own currency. As for macroeconomic costs, they depend on the nature of the shocks faced by the economy. A vast empirical literature has aimed at identifying the nature of the shocks faced by individual countries in would-be monetary areas: Europe, East Asia, Latin America, etc. Various methodologies can be used, ranging from calculating business-cycle correlations across countries to more sophisticated econometric methods aimed at measuring shock asymmetry, or more structural measures such as openness, industry diversification, or intra-industry trade (see box 5.13).

Box 5.13 Identifying Asymmetric Shocks

Descriptive Methods

To identify the need for a flexible exchange rate, a simple way is to calculate observed *real* exchange-rate volatility: Whatever the exchange-rate regime, a volatile real exchange rate means that there is a need for relative price adjustments, which are easier if the nominal exchange rate is allowed to adjust. The problem with this method is that real exchange-rate volatility can be driven by speculative movements in a floating regime, whereas in a fixed exchange rate, nominal exchange-rate rigidity can be an obstacle to real exchange-rate adjustment. Hence, this first measure of shock asymmetry is not free from the exchange-rate regime itself.

A more popular measure of shock asymmetry relies on cross-country correlation of variables such as industrial production, GDP or employment: The higher the cross-country correlation, the less need for real exchange-rate adjustment. Figure B5.13.1 shows the correlation of real GDP-per-capita growth rates between a number of euro and non-euro countries and the euro area aggregate, for two periods: 1971–89, and 1993–2003.[a] The correlation is generally high over the second period (around 70–80%), with the notable exception of Greece, where the correlation drops from +0.48 in the first sub-period to –0.20 in the second one. The reverse pattern applies to Finland and Ireland, which show a strong convergence on the euro area business cycle. Interestingly, such convergence can also be observed in Denmark, Sweden, and the UK, which have not yet decided to join the euro area. As for non-European countries, a convergence with the euro area can be observed in Canada, but not in the US, or, especially, Japan.

The problem with this method is that a high correlation between two countries can come either from symmetric shocks to output or from strong policy reactions, including independent monetary-policy reactions, to asymmetric shocks. In order to measure shock asymmetry, econometric estimations must be carried out.

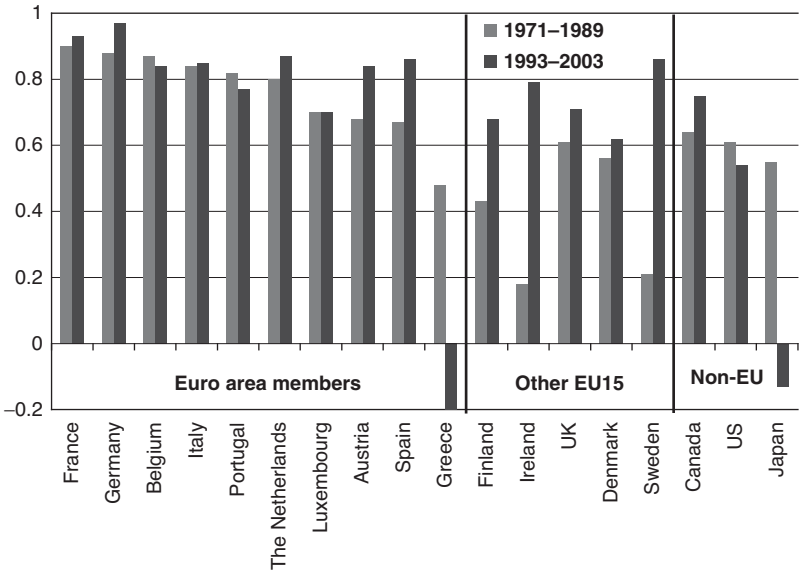


Figure B5.13.1 Correlation of GDP per capita growth with the euro area. Source: Giannone and Reichlin (2006).

Econometric Methods

In econometric terms, exogenous shocks can be identified using the residuals of estimated equations. In order to account for interdependence between variables and/or countries, VAR methodologies have been developed. Blanchard and Quah (1989) suggest estimating a VAR model with two variables (the logarithm of output, and the logarithm of the price level), and identifying demand shocks and supply shocks by applying an identification matrix to the residuals of the VAR model. The identification matrix is based on a normalization of the shocks, on the assumption that demand and supply shocks are orthogonal, and on the additional hypothesis that demand shocks have no long-run impact on output, consistent with the aggregate-supply–aggregate-demand model (see chapter 1). This exercise can be performed for each country successively, and cross-country shock correlations can be derived from it. Using this methodology, Bayoumi and Eichengreen (1994) found that only a core group of European countries exhibited a correlation of shocks comparable to the correlation across US regions. However, the result could be attributable to the exchange-rate regime itself, since only a core group around Germany experienced stable exchange rates over the sample period. Another problem is that the shocks originating in foreign countries are wrongly identified as domestic shocks in this closed-economy setting.

Another avenue consists in estimating a two-country VAR model. For instance, Giannone and Reichlin (2006) estimate a VAR model with

two variables: Real GDP per capita in each euro area country, successively, and average real GDP per capita in the euro area, over 1970–2003. Then, they are able to measure the percentage of GDP per capita deviations due to country-specific shocks in the short run and in the longer run. Figure B5.13.2 reproduces their results. The share of country-specific shocks in euro-member output fluctuations is generally small, especially after five years. However, the same outliers as in figure B5.13.1 emerge: In Ireland, Finland, and Greece, more than 80% of output fluctuations come from country-specific shocks. Note, however, that the estimation is performed over the entire 1970–2003 period, whereas figure B5.13.1 evidences a convergence of Finland and Ireland after 1993.

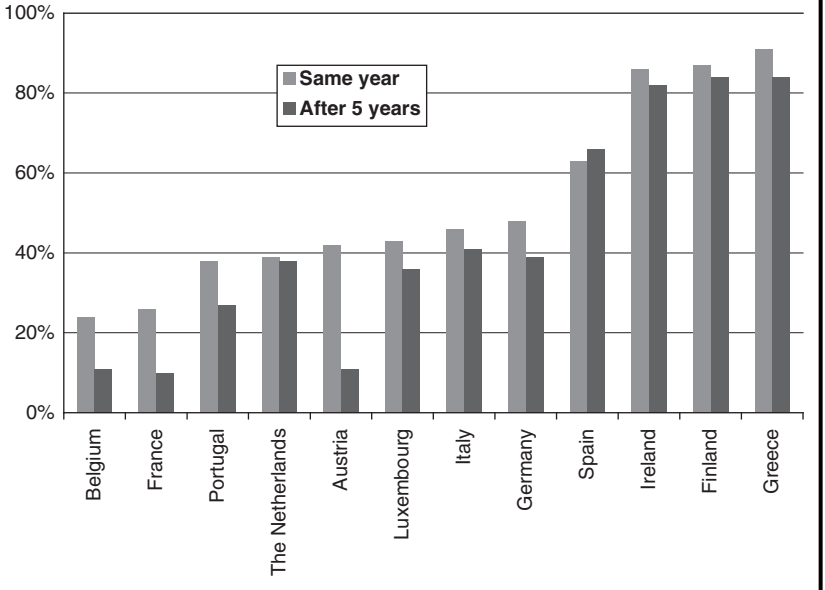


Figure B5.13.2 Share of country-specific shocks in explaining output fluctuations. Source: Giannone and Reichlin (2006).

Measure of Trade Specialization

The difficulty with the above methods is that they are backward-looking. Another way to measure the scope of asymmetric shocks is to look at the extent of industry specialization. A synthetic measure of specialization is given by the Herfindhal index:

$$H = \sum_{i=1}^{n} \left(\frac{Y_i}{Y} \right)^2$$

where Y_i is output in industry i, Y is total output and n is the number of industries. $H = 1$ if the country is specialized in a single industry and is lower the more diversified the economy. An alternative measure of

specialization is the extent of inter-industry trade. For instance, the Finger–Kreinin index can be calculated as follows:

$$FK = \sum_{i=1}^{n} Min\left(\frac{X_i}{X}; \frac{M_i}{M}\right)$$

where X_i and M_i denote exports and imports of product i, respectively. FK varies from zero (only inter-industry trade) to one (only intra-industry trade). Another measure of intra-industry trade is the Grubel–Llyod index:

$$GL = \sum_{i=1}^{n} \left(\frac{X_i + M_i}{X + M}\right)\left(\frac{(X_i + M_i) - |X_i - M_i|}{X_i + M_i}\right)$$

This index also varies from zero (only inter-industry trade) to one (only intra-industry trade). The advantage of relying on trade data is that it is available at a detailed level, allowing for a precise measure of specialization. For instance, a high reliance on the chemical industry may not have the same meaning depending on the range of chemical items produced and traded (mineral, pharmaceutical . . .). Figure B5.13.3 ranks EU countries depending on the share of intra-industry trade, measured through the Grubel-Llyod index. Again, Greece stands out, followed by Finland and Ireland.

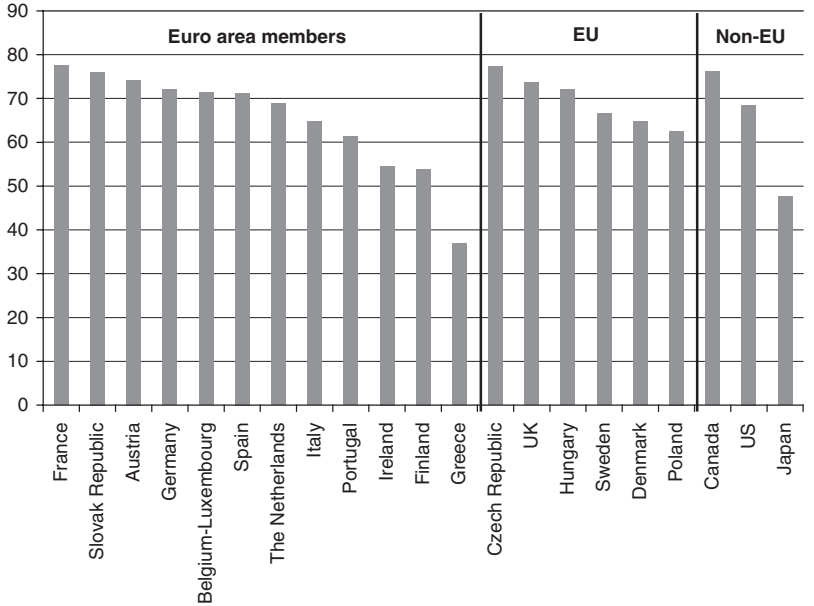


Figure B5.13.3 Grubel-Lloyd index of intra-industry trade, in percent. Source: OECD (2002).

[a]The 1990–92 period is excluded due to German reunification.

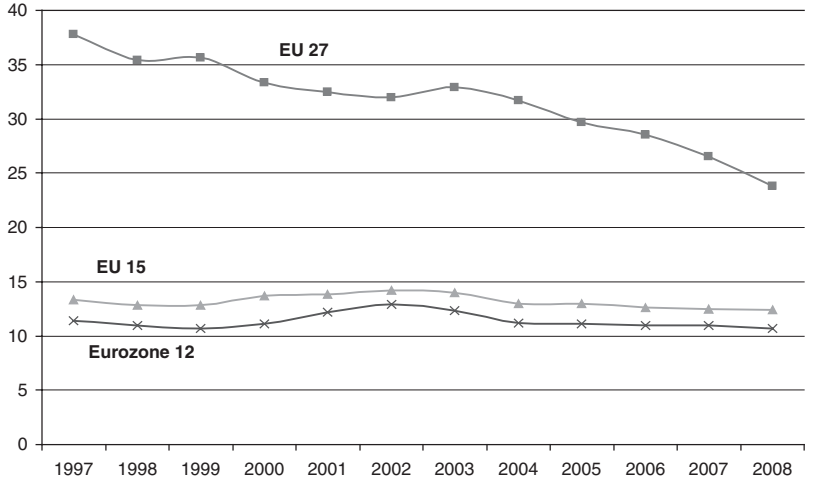


Figure 5.19 Price convergence in the EU, 1991–2008.
Source: Eurostat, structural indicators.

Reading: Coefficient of variation of comparative price levels of final consumption by private households, including indirect taxes, in %.

In practice, the theory of optimum currency areas has seldom been used to decide whether or not to proceed to monetary union. In the European case, monetary union has been a final goal for several decades as a way to promote political integration and to coordinate better macroeconomic policies by avoiding beggar-thy-neighbor policies. Policymakers in Continental Europe were also convinced that monetary union was a necessary complement to the single market. Irrevocably fixed exchange rates and lower conversion costs would enhance price transparency and favor trans-national investments. Figure 5.19 shows that, indeed, price dispersion has been lower in the euro area than in the EU-25.

However, prices did not converge faster within the euro area than within the EU-15 between 1997 and 2008. In 2008 price dispersion (a coefficient of variation of 11% for consumption prices) was still higher than in the US.[34] This shows that monetary union can only complement the deregulation of goods and services markets. In the euro area, some markets, especially in service sectors, have remained heavily regulated until the mid-2000s, preventing price convergence. The European service directive will likely be a stronger vector of price convergence than the mere single currency.

Other regions in the world have started to discuss monetary integration, in the form of monetary unions (East Asia, West Africa, the Caribbean, Gulf countries) or in the form of "dollarization" (NAFTA, New-Zealand ...).

34. See Ilzkovitz et al. (2007). Nevertheless, the euro has reduced price discrimination of *exporters* within the euro area (see Fontagné et al., 2009).

As illustrated in box 5.14 in the case of West Africa, the theories developed in section 5.2 are generally not at the core of the discussions.

Box 5.14 Monetary Borders in West Africa

There are two monetary unions in sub-Saharan Africa: The WAEMU (*West African Economic and Monetary Union**, in French the UEMOA), and the CAEMC (*Central Africa Economic and Monetary Community**, in French the CEMAC). They share the same currency, the *CFA franc* and they constitute the "CFA franc area" together with a small Indian Ocean archipelago, the Comoros. Both areas are fully fledged monetary unions, with single monetary policies and foreign reserves centralized within two central banks, the Dakar-based Banque centrale des États d'Afrique de l'Ouest (BCEAO) and the Banque des États d'Afrique centrale (BEAC) based in Yaoundé. Convertibility of the CFA franc at a fixed rate vis-à-vis the euro is guaranteed by the French government under a mechanism comparable to a currency board (convertibility is backed by a euro-denominated deposit at the French Treasury).

Even though their currencies are pegged, the underlying economic structures of the two monetary unions are in stark contrast: Western African countries are oil importers while several Central African countries are oil exporters. Also, the relevance of monetary borders is increasingly being questioned. In 2000, six non-WAEMU-member West African countries stated their intention to establish a monetary union in 2003, then have it merge with WAEMU and make monetary borders coincide with the Economic Community of the West African States (ECOWAS), a wider grouping which includes WAEMU. Convergence criteria were defined, and a West African Monetary Institute was established in Accra (Ghana) to organize multilateral monitoring and prepare for monetary unification. In April 2002, the West African Monetary Area (ZMOA) was created between five countries (Gambia, Ghana, Guinea, Nigeria, Sierra Leone), with an exchange-rate mechanism limiting the fluctuations of each currency to $+/-$ 15% in relation to the US dollar. Disappointing results as regards convergence led governments to postpone monetary union.

This new monetary dynamics in West Africa raises the question of optimal monetary borders in sub-Saharan Africa. The question is particularly difficult for low-income, very specialized countries which do not trade much with each other. Empirical analysis tends to conclude that the Economic Community is not an optimum currency area: A "core group" of countries can be identified within WAEMU based on optimal currency area criteria, but the CAEMC is very heterogeneous and Nigeria, a large oil exporter, stands as a special case (Masson and Pattillo, 2001, 2005; Bénassy-Quéré and Coupet, 2005).

A rare example of a country having explicitly used theory-based analysis to assess its participation in a monetary union, both in the micro and the macro dimensions, is the UK's assessment of the desirability of euro membership—the "five tests" defined by then-Chancellor Gordon Brown in 1997 (box 5.15).

Box 5.15 Five Tests to Join the Euro

The UK is one of the two countries (together with Denmark) benefiting from an "opt-out" from the euro in the Maastricht Treaty. In 1997, Tony Blair's new government outlined five economic "tests" to decide on UK entry in the euro, in addition to the Maastricht criteria that would be used by the European Commission and ECB to assess convergence. The questions were the following: (i) Are the UK business cycle and economic structure compatible with a single, euro-area-wide interest rate? (ii) Would the UK economy remain sufficiently flexible in case of adverse economic events? (iii) Would euro entry foster investment in the UK? (iv) How would it impact the competitiveness of the UK financial sector? (v) Would euro entry be favorable to growth? Her Majesty's Treasury then built models and commissioned reports to academics. In June 2003, the conclusions went as follows:

> Overall the Treasury assessment is that since 1997 the UK has made real progress toward meeting the five economic tests. But, on balance, though the potential benefits of increased investment, trade, a boost to financial services, growth and jobs are clear, we cannot at this point in time conclude that there is sustainable and durable convergence or sufficient flexibility to cope with any potential difficulties within the euro area. So, despite the risks and costs from delaying the benefits of joining, a clear and unambiguous case for UK membership of EMU has not at the present time been made and a decision to join now would not be in the national economic interest.
>
> HM Treasury, UK Membership of the Single Currency: An Assessment of the Five Economic Tests, June 2003, p. 6

c) Credible exchange-rate regimes

While economists generally put forward the shock-absorbing properties of an exchange-rate regime, governments rather emphasize its contribution to the credibility of economic policy.

As discussed in chapter 4, monetary policy is not always credible. This can be due to a track record of economic-policy mistakes (e.g., in Latin America in the 1980s) or to the absence of a policy track record at all (e.g., in countries in transition such as Eastern European countries in the 1990s, or in newly established countries such as Timor-Leste or Kosovo). External anchoring

Table 5.5
Inflation and growth performance under various exchange-rate regimes

	CPI inflation	GDP growth
Pegged	8.4%	1.4%
Intermediate	11.6%	2.1%
Floating	15.2%	1.7%

Source: Gosh et al. (1997), based on 36 countries over 1960–90.

can then be used to import credibility. By announcing that the exchange rate will not move, the central bank achieves a double end: It incites firms and employees to ask for moderate price and wage increases, and it ties its own hands to make monetary surprises impossible. This justifies *ex post* the fixed exchange rate. In short, it helps the economy settle on the right equilibrium.

Although the causality of such a relationship is necessarily difficult to fully establish, a number of case studies seem to confirm that an exchange-rate peg can curb inflation efficiently, while the impact on growth is unclear (see Table 5.5).

Figure 5.20 illustrates the spectacular example of Argentina in the early 1990s, after a currency board was put in place. A more modest case is that of the French policy of "franc fort" in the late 1980s and early 1990s, when policy authorities strove to keep the peg to the German mark despite rising unemployment, and successfully brought the inflation rate below the German level in the 1990s. However, currency crises of the late 1990s have proved that external anchoring cannot be a durable substitute for internal credibility. Even a hard peg is vulnerable, as shown again in the case of Argentina when the currency board had to be abandoned at the end of 2001 in the midst of a political and social crisis.

Credibility can be enhanced by posting a clear long-term strategy. In Europe, political commitment to monetary union helped stabilize currencies in the second half of the 1990s, and the prospect of eventually joining the euro has played the same role for Eastern European currencies. Under the Maastricht Treaty, these countries are to be accepted into the euro area after participating for two years in the *European Exchange Rate Mechanism** (ERM), which requires their nominal exchange rate to fluctuate at most by $\pm$ 15% around central rates vis-à-vis the euro. The ECB is committed to defending the central rate by buying or selling currency against the euro at the limit.[35] This creates expectations of nominal exchange-rate stability. Although, this

35. The ECB is committed insofar as that does not contradict its primary objective of maintaining price stability. The ECB could refuse to intervene to support an exchange-rate-mechanism currency if this implied creating too many euros in exchange. This is very unlikely, given the small size of the markets for these currencies.

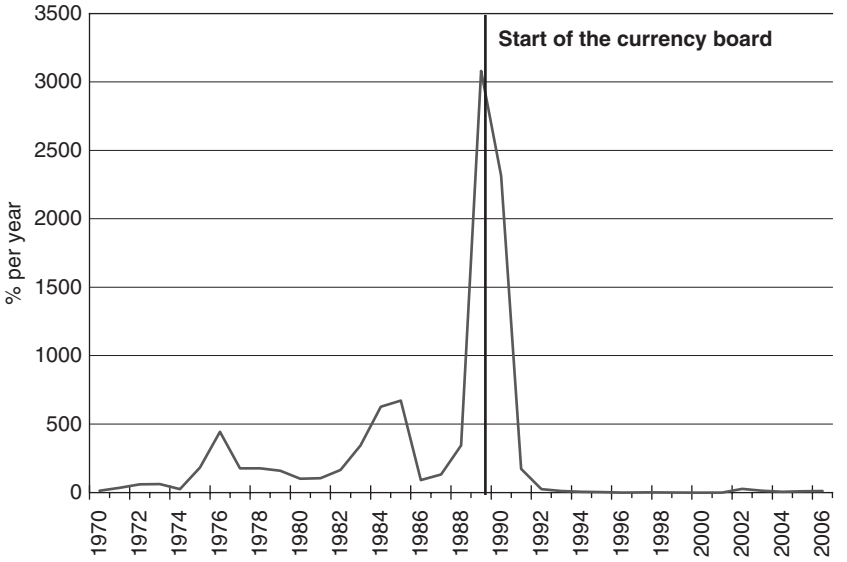


Figure 5.20 Disinflation in Argentina: Inflation rate, 1970–2006.
Source: World Bank, *World Development Indicators*.

has proved workable for smaller economies, larger countries like Hungary or Poland have chosen to retain exchange-rate flexibility and not to join the ERM, at the cost of postponing their accession to the euro.

On the whole, the choice of an exchange-rate regime depends on a number of criteria that can have different weights depending on the country and period in time. As Jeffrey Frankel wrote, "no single exchange rate regime is right for all countries and at all times" (Frankel, 1999).

5.3.2 Managing floating exchange rates

In a floating exchange-rate regime, the exchange rate moves in order to equalize the supply and demand for the currency. Why should governments try to influence this process? There are three reasons. First, they can aim at bringing the exchange rate back toward what they regard as the *economic fundamental**, i.e., a value they deem consistent with macroeconomic equilibrium (the equilibrium exchange rate of section 5.2). Given the uncertainty surrounding estimates of equilibrium exchange rates, there is much credibility to lose in the exercise and governments usually step in only when market rates have departed very substantially from equilibrium values. More ambitiously, they can aim at using the exchange rate as a policy instrument to bring the whole economy to a different equilibrium point. This is typically the reason behind devaluations within fixed exchange-rate regimes. Thirdly, without reference to a particular level of the exchange rate, they can

aim at reducing its volatility because they regard it as costly for economic welfare.

Governments affect the exchange rate by buying or selling foreign currency on the marketplace in a *foreign-exchange intervention**. One can distinguish three groups of countries depending on their attitude. A first group of countries with managed float regimes includes most emerging market economies. These countries monitor closely their currency by intervening often, sometimes daily, on foreign-exchange markets. A second group of countries with floating-rate regimes, including the US, the UK, and the euro area, intervene very seldom (the Fed did not intervene between August 1995 and September 2000, and between September 2000 and 2009), but retain the option of doing so. At periods, Japan has been a member of the former or of the latter group. Finally, Australia is a rare example of a country that refrains altogether from intervening. New Zealand has long been part of the third category but eventually intervened in June 2007 to sell its currency after its value had soared against the Japanese yen due to *carry trades**, i.e., investors borrowing in yen to invest in New Zealand dollars, to cash in the interest-rate differential between short-term deposits denominated in the two currencies.

a) Who is in charge?

The institution responsible for monitoring exchange-rate regimes and exchange-rate policies, in relation to global current account imbalances, is the International Monetary Fund. The Articles of Agreement of the IMF were modified in 1976 to legalize floating exchange rates, taking stock of the suspension of the gold convertibility of the US dollar and of the devaluation of the dollar in 1971. According to article IV of the IMF Articles of Agreement, "Each member shall notify the Fund . . . of the exchange rate arrangements it intends to apply," and the Fund ". . . shall exercise firm surveillance over the exchange rate policies of members." This was detailed in a 1977 decision on surveillance over exchange-rate policies of member countries, which was not changed until June 2007, when it was decided to account for monetary unions, but also for spillovers of the exchange-rate regime of a given country to other countries and to financial markets at large. At the insistence of the US, the revised decision also stresses the need to avoid "exchange-rate manipulation", a clear allusion to China's exchange-rate policy.[36]

However, there is not much the Fund can do to "exercise firm surveillance over exchange-rate policies." It cannot twist the arms of its members, except for those seeking financial assistance, in which case foreign-exchange policy can be included in the conditions attached to the loan. As a result, the Fund has shied away from designating misaligned currencies.

36. *Review of the 1977 Decision on Surveillance over Exchange Rate Policies*, available on the IMF Web site.

Before the G20 was chosen in 2009 as the "premier forum for international economic cooperation," exchange rates between major economies, and more generally in the global economy, were mostly discussed by G7 ministers of finance and central bank governors.[37] Statements released after G7 finance-minister meetings would typically include a section on exchange rates, usually written in arcane language (a wording such as: "We agreed to monitor the exchange-rate situation carefully and act as appropriate" is meant to send the signal that the G7 is ready to intervene on foreign-exchange markets). In the 2000s, this has increasingly gone along with bilateral policy dialogue with large emerging-market economies, in particular with China. In 2009, G20 leaders established the G20 "framework for strong, sustainable and balanced growth" to compare and mutually assess their growth strategies, based on IMF analysis. Looking forward, the G20 could play a greater role in exchange-rate discussion.

In all countries, choosing the exchange-rate *regime* is a political decision, responsibility for which rests on the government. But depending on countries, exchange-rate *policy* is decided either by the central bank or the ministry of finance. In the US, Japan, UK, and Canada, the exchange-rate policy is solely the responsibility of the finance minister, with the central bank acting as his or her agent. In the US federal framework, the Federal Reserve Bank of New York undertakes all market operations, including foreign-exchange interventions.

In pre-EMU Europe, the German central bank, the Bundesbank, had acquired broad autonomy as regards exchange-rate policy, in view of the overriding priority given to price stability. In several instances, the Bundesbank refused to stabilize the external value of the deutschemark on the grounds that excessive money creation to purchase foreign currencies would jeopardize internal monetary stability.[38] This was the case in May 1971, when the US dollar began to depreciate after the US current-account deficit had suddenly widened, and again in September 1992, when the pound sterling and the Italian lira came under attack in the European exchange-rate mechanism. Since 1999, the euro area has remained close to the German tradition. Whether the EU Treaty provides for an effective exchange-rate policy is unclear. Unlike the US, UK, or Japan, the responsibility for exchange-rate policy is legally shared between ministers of finance and the ECB, and it is practically in the hands of the latter. "I am Mr. Euro," Wim Duisenberg, the ECB's first president, once famously said. The ECB and national central banks manage foreign-exchange reserves without government interference and are solely responsible for market interventions (unlike the US Federal Reserve system, all national central banks then have the possibility of intervening, under the leadership of the ECB). Ministers can only negotiate formal monetary

37. For more on the G7 and G20, see chapter 2.

38. This was stated as a principle in a secret letter sent by Bundesbank president Otmar Emminger to the German minister of finance at the time of the creation of the European Monetary System.

agreements with third countries[39] and formulate "general orientations for exchange-rate policy" on a proposal from the Commission or the ECB, and provided these do not endanger price stability (Lisbon Treaty, article 2). Hence, any deliberate weakening of the euro to prop up output is ruled out, except in total absence of inflationary pressures. Moreover, such "general orientations" have never been issued.

b) The intervention toolbox

As presented in box 5.6, there are three ways for monetary authorities to influence (or try to influence) the exchange rate: Official interventions, interest-rate variations, and communication. When they want to avoid a depreciation of the domestic currency, monetary authorities can: (i) Sell foreign currencies on the international money market; (ii) raise the domestic interest rate (to attract foreign investors); or (iii) publicize some private information on exchange-rate fundamentals, or simply on their willingness to defend a certain level of the exchange rate. If investors are risk-neutral, then the exchange rate follows the uncovered interest-rate parity. In this case, official interventions cannot affect the exchange rate unless they are accompanied by monetary policy and/or communication. If investors are risk-averse, then an intervention can be powerful, as evidenced by the case of East-Asian countries in the 2000s (see figure 5.21).

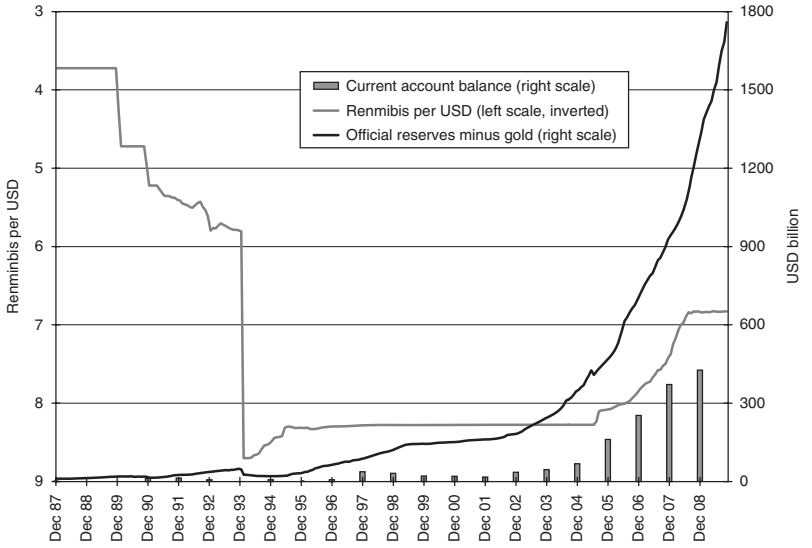


Figure 5.21 The success of foreign-exchange interventions in China, 1994–2007. Source: International Monetary Fund IFS database.

39. This provision was used only to negotiate on euro adoption by small, non-EU countries such as the Republic of San Marino.

As already discussed, purchasing foreign currency implies creating money. To offset this impact on monetary aggregates, the central bank can sterilize its interventions by reducing by the same amount its claims on the domestic economy, e.g., by selling securities to domestic agents. These can be government or corporate securities taken from the central bank's portfolio, or bills issued by the bank for that specific purpose and called *sterilization bonds**. Monetary policy is then immunized from the impact of exchange-rate policy. In the early 2000s, changes in net domestic credit and nonmonetary liabilities have offset between 85% and 95% of changes in net foreign assets in India, Korea, Malaysia, Singapore, and Taiwan, and over 70% and 60% in China and Russia (Mohanty and Turner, 2006).

Sterilized interventions are less effective than nonsterilized ones since, by definition, they do not change the relative supply of money (box 5.16). Moreover, in countries with a weak currency and/or a high risk premium, yields on domestic securities are much above yields on foreign securities. The cost of carrying foreign assets weighs on the budget, and a weakening fiscal position can in turn generate indirect inflationary pressures. In the 2000s, the budgetary cost of sterilization in emerging countries was nevertheless limited thanks to low risk premia, reaching at most 1% of GDP in Indonesia based on market interest rates (Mohanty and Turner, 2006). The People's Bank of China was even earning a positive carry, the one-year Chinese interest rate being less than half the comparable US Treasury rate.

Sterilized interventions are all the more effective insofar as their amount is commensurate to the turnover of foreign-exchange markets. That is the case in some emerging markets but not for the main currencies, for which the volume of transactions reaches several hundred billion dollars per day. However, interventions still have a signaling effect and help alter expectations.

Box 5.16 Are Foreign Exchange Interventions Effective?

One usually identifies three channels through which foreign-exchange interventions can impact the exchange rate (Dominguez and Frankel, 1993).

- *The monetary channel** Suppose that a central bank opposes the appreciation of its currency. It purchases foreign currency in exchange for central-bank money. This increases simultaneously the assets of the central bank, of which official reserves are part, and its liabilities, through currency in circulation. The increase of money supply is magnified by a multiplier effect through the banking system. The overall monetary expansion that results from it lowers the interest rate. The lower yield on domestic assets discourages capital inflows and halts the appreciation of the currency. In fact, the

foreign-exchange intervention amounts to a loosening of monetary policy. This channel is important in emerging market economies, where financial markets are not developed (hence the scope for open market operations is limited domestically), and where the budgetary cost of sterilization is high in relation to GDP.

- *The portfolio channel** If assets denominated in various currencies are imperfect substitutes, then a change in the relative supply of domestic and foreign assets will affect their relative price, i.e., the exchange rate. Again, consider a central bank that opposes the appreciation of its currency. To this end, it purchases foreign currency, which increases the relative supply of domestic currency and causes the exchange rate to drop. The effectiveness of this channel relies on imperfect capital mobility and/or investors' risk-aversion (see box 5.6). Empirical studies tend to downplay this channel when capital is mobile, since the amounts under consideration are insufficient to alter the balance of the foreign-exchange market. For the same reasons, capital controls can enhance the effectiveness of the portfolio channel. The typical example is China in the early 2000s, when massive official interventions could halt the appreciation of the renminbi in spite of the doubling of the current-account surplus and massive foreign direct investment inflows (figure 5.21).
- *The signaling channel** When intervening, the central bank changes the structure of its own balance sheet in a way that aligns its incentives with the signal that is being sent to the markets. For example, if it sells foreign currency in exchange for domestic currency, it is in its own interest that the domestic currency does not depreciate, otherwise it will take a loss on this operation. By intervening, it reveals either that it holds private information on economic fundamentals that backs a scenario of appreciation, or that the government is politically committed to having this scenario materialize. This signaling effect is particularly important when market participants do not have a precise idea of the equilibrium value of the exchange rate.

Empirical studies (Frenkel et al., 2001; Sarno and Taylor, 2001) cast a doubt on the effectiveness of sterilized interventions. Based on historical data, they find that the impact is counter-intuitive: The domestic currency depreciates following foreign currency sales. One possible explanation is that the central bank sells foreign currency when its domestic currency depreciates. However, empirical work confirms that in general, interventions have weak effects on the exchange rate. Similarly, interventions seem to increase rather than reduce exchange-rate volatility. Fratzcher (2004) finds that public statements on the exchange rate by

monetary policy committee members in the US, euro area, and Japan do affect exchange rates in the short term and that they reduce their volatility. Econometric studies should be taken with a grain of salt since they isolate interventions from their general economic and political context. On the whole, there seems to be a consensus on the fact that interventions are effective when they support decisions or public announcements related to exchange-rate fundamentals, i.e., monetary and fiscal policy, but this is only partially backed by empirical studies.

c) The dollar, the euro, and the yen

The exchange-rate policies of the three main currency areas deserve specific attention. The US dollar's role as the main international currency allows a certain neglect with respect to the external value of the currency. As noted in section 5.2, a lower dollar does not necessarily raise the cost of servicing the external debt of the nation, and dollar securities easily find their way into foreign portfolios in a world where the dollar is the main reserve of value. Symmetrically, a strong dollar raises the purchasing power of US residents and makes it easier to finance the foreign deficit. A simple, arithmetic point is worth recalling. In a world with n currencies, there are only $n - 1$ independent exchange rates. If $n - 1$ countries manage their exchange rates, the remaining one is entirely determined. Alternatively, if all n countries seek to manage their exchange rates, they have to coordinate in some way or they will clash. This is sometimes called the *nth country problem**. In the words of Peter Kenen, " ... the United States did not try to pursue an independent exchange-rate policy in the first decades following World War II. Because of its great economic strength and comparative self-sufficiency, it was content to act as the *n*th country in the system." (Kenen, 2000, p.277). When it comes to the US exchange-rate policy in the post-Bretton-Woods era, *benign neglect** has been the name of the game.

At times, the US has sought to steer its exchange rate when it was considered too weak (adding to inflationary pressures) or too strong (switching expenditures toward goods produced abroad). This has been undertaken in a coordinated way with other industrialized countries, producing the *Plaza agreement** of 25 September 1985, when G5 members committed to lowering the value of the dollar, and the *Louvre agreement** of 22 February 1987, when G6 members temporarily agreed on a system of exchange-rate *target zones**, i.e., fluctuation bands for bilateral nominal exchange rates which would be protected by central bank intervention at the margin of the bands. The US also joined the G7 on 15 August 1995 to weaken the yen, and on 25 September 2000 to support the euro.

From a political economy standpoint, the main constituency supporting currency intervention is made up of the representatives of states exposed to foreign competition, and particularly to industry relocation outside

the US. Most economists agree that competition of low-wage, labor-intensive countries is a feature of globalization and that exchange rates play only a limited role in compounding its effects. However, it has been difficult for US administrations to resist using exchange-rate intervention to deflect protectionist pressures. In the late 1990s, the Clinton administration managed to maintain low interest rates through a controlled communication on the "strong dollar" policy (box 5.17). In contrast, in the 2000s, the Bush administration let the dollar depreciate as a reaction to the current-account deficit and to perceived job losses to China.

Box 5.17 The Strong Dollar According to Robert Rubin

Robert Rubin, a former Goldman Sachs executive and Bill Clinton's Secretary of the Treasury from 1995 to 1999, linked his name to the strong dollar policy. In his memoirs, he gives an account of his principles and methods:

> A strong currency means that American consumers and businesses can buy imported goods and services more cheaply and that inflation and interest rates will tend to be lower. It also exerts pressure on American industry to increase productivity and competitiveness. These benefits can feed on themselves as foreign capital flows in more easily because of greater confidence in our currency. A weak dollar would have the contrary effects.
>
> R. Rubin and J. Weiseberg (2003), p. 182

> Because of Treasury's ability to buy and sell currencies for the purpose of affecting exchange rates, the markets would response to almost anything I said that seemed to make intervention more or less likely. Affecting exchange rates unintentionally would make me look undisciplined and unsophisticated. . . . whatever my views were about whether the dollar at any given moment was too strong or too weak relative to economic fundamentals, I virtually always said exactly the same thing: *"A strong dollar is in our national interest."* . . . The slightest shading, such as going from *"I believe a strong dollar is in our national interest"* to *"I believe it's in our national interest to maintain a strong dollar"* could have market effects, even if no change in view was meant. [. . .] my saying *"A strong dollar is in our national interest, and we have had a strong dollar for some time now"* created great excitement at a press briefing, as it was construed to mean that we wouldn't mind seeing the dollar remain strong but soften somewhat.
>
> R. Rubin and J. Weiseberg (2003), pp. 182 and 184

The attitude of the euro area is comparable in practice to that of the US despite its different constitutional framework. There is a consensus among European institutions that market intervention is desirable only in exceptional circumstances, in case of a clear misalignment of the exchange rate

with respect to its fundamental value (as defined in section 5.2) or of an excessive volatility. Apparently, this has occurred only once since the introduction of the euro: In 2000, the value of the euro was 30% lower than when it was created in 1999 and the ECB asked other G7 central banks for help, leading to a joint intervention on 25 September. As already mentioned, the "general orientations of exchange rate policy" envisaged in the EU Treaty were never used. One cannot exclude the possibility of a conflict between finance ministers and the ECB in a case where inflationary pressures and a strong exchange rate would coexist. The ECB would then raise interest rates to fight inflation, making the exchange rate appreciate further. Since the Treaty gives priority to price stability, ministers could not issue general orientations to force the ECB to sell euros, and the monetary and political arms of the Union would be at odds.

Japan has intervened much more frequently than the US or Europe in foreign-exchange markets to oppose the appreciation of the yen resulting from the current-account surplus. Interventions reached unprecedented amounts in the early 2000s. The Bank of Japan sold 1284 billion yen (around 10 billion dollars equivalent) in one trading day on 10 December 2003. This strategy was generally considered to be successful (box 5.18), partly because interventions were massive, were not sterilized and were adding to the effects of monetary policy. Japan stopped intervening in March 2004, as GDP growth had materialized: A recovering domestic demand was expected to dampen the current-account surplus and make currency appreciation less necessary. In addition, nonsterilized interventions to sell yen would have been in contradiction with monetary-policy tightening, while sterilized interventions would have been less efficient (see above)

Did floating exchange rates fulfill their roles as macroeconomic shock absorbers in the three main economic areas? For the euro area, this has not been the case. Positive shocks to aggregate demand have not led to a real exchange-rate appreciation, as the equilibrium exchange-rate approach would predict. Figure 5.22 plots the real effective exchange rate against the output gap for the US and the euro area. Over the period 1996–2000, the output gap gradually turned positive in the euro area and the euro depreciated against all currencies (remember that G7 central banks intervened in September 2000 to reverse this trend). Over the period 2000–04, the output gap went back into the red and the euro appreciated against its partners. Graphically, up to 2007, the real effective exchange-rate path was confined to the north-west and south-east corners of figure 5.22. In short, between 1996 and 2006, the exchange rate has played a procyclical role, magnifying fluctuations of aggregate demand instead of dampening them. This has been less the case in the US, where the exchange rate has generally played more of a counter-cyclical, shock-absorbing role.

Why did the euro exchange-rate play such a procyclical role? A first answer has to do with monetary-policy reaction functions on both sides of the Atlantic. As taught by Mundell and Fleming, in a world of mobile capital

Box 5.18 Japanese Foreign-Exchange Interventions in 2003

Japan is the developed country where exchange-rate policy has been the most active. The current-account surplus has pushed the yen upward, hampering the price competitiveness of Japanese firms. Until 2004, the Ministry of Finance bought dollars (and to a lesser extent euros) on a daily basis on the foreign-exchange market. To finance these purchases, it issued so-called *financing bonds* subscribed by the Bank of Japan or by commercial banks. The impact of these interventions on money supply was not sterilized because they were aligned with the quantitative monetary-expansion policy followed by the central bank. In a more restrictive monetary context, the Bank of Japan would have sterilized interventions by selling securities on the open market, thereby withdrawing money from the banking system.

The dates and amounts of intervention have been disclosed by the ministry of finance. In 2003, it intervened 82 times and sold a total of 20425 billion yen, of which 178 billion were against euros; 1284 billion yen were sold on 10 December. As can be seen in figure B5.18.1, interventions did stabilize the dollar/yen parity around 120 yen/dollar until the autumn of 2003, but they could do nothing against the brutal appreciation toward 107 yen/dollar that took place in September.

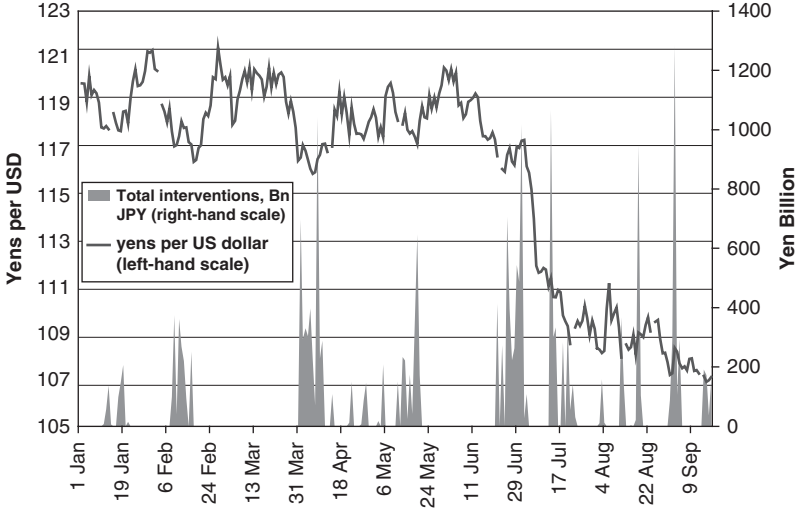


Figure B5.18.1 Japan's foreign-exchange interventions and the dollar/yen exchange rate, Jan–Sep 2003.
Source: Japanese Ministry of Finance, Reuters, and authors' calculations.

Note: A decrease indicates an appreciation of the yen.

Figure B5.18.2 shows the variation of the dollar/yen parity on days when interventions took place. The impact is limited: All things being equal, 100 billion yen sold on the market cause a 0.07% depreciation of the end-of-day exchange rate, and only large-scale interventions seem to have been effective.

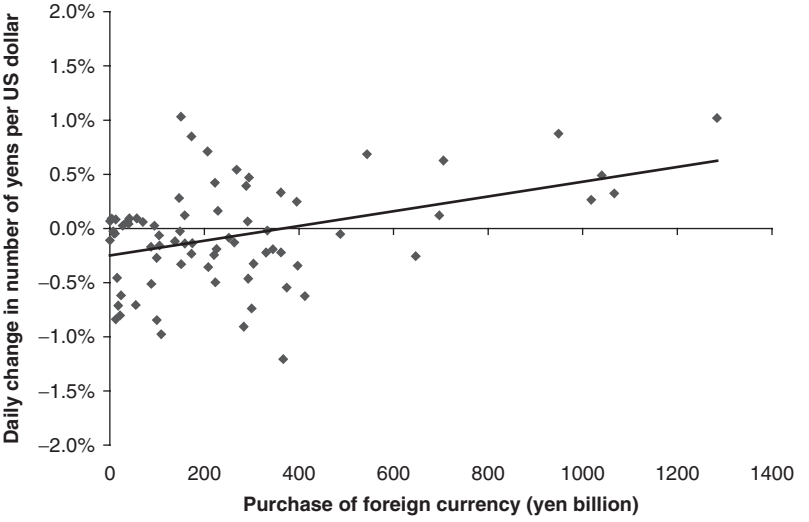


Figure B5.18.2 Interventions and exchange-rate variation.
Source: Japanese Ministry of Finance, Reuters, and authors' calculations.

and floating exchange rates, the impact of monetary-policy decisions on output is magnified by exchange-rate movements. Indeed, over the business cycle, the US Federal Reserve reacted more swiftly to demand shocks than the ECB (see chapter 4). Another answer has to do with the economic policy frameworks of the euro area and the US. As already discussed, exchange-rate policy principles in the euro area are minimal and the exchange rate has been considered a residual variable rather than a policy instrument. This has not been the case in the US where the Treasury conducted successively (and successfully) a strong dollar policy, then a policy of "benign depreciation."

Figure 5.22 illustrates a paradox of today's international monetary arrangement. While the business cycles are increasingly synchronized due to international trade in goods and services and to integration of financial markets (Bordo and Hebling, 2003), exchange-rate fluctuations remain wide. This is because shocks increasingly originate from the financial account. Shifts in investor preferences, foreign direct investments, and short-term speculation on the foreign-exchange market all result in exchange-rate movements that shift output from one area to another. In some instances, this justifies direct currency intervention.

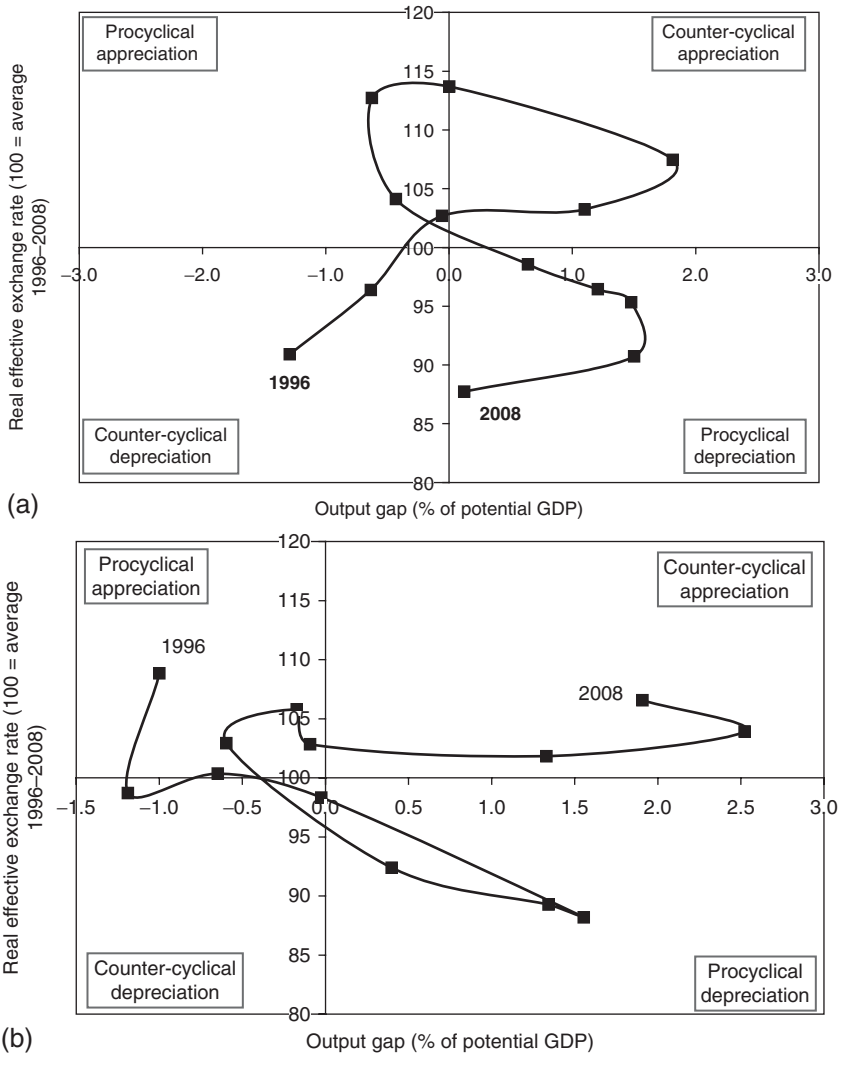


Figure 5.22 Output gaps and real effective exchange rates in the US and the euro area. a) US, b) euro area, 1996–2008.
Source: *OECD Economic Outlook*, November 2008 and April 2009.

5.3.3 The future of the international monetary system

The international monetary system has changed radically since the Bretton Woods conference of 1944. Since 1972, industrialized countries have lived with floating exchange rates, and since the 1990s, they have lived with freely mobile capital. This was not anticipated in Bretton Woods, as international adjustment was expected to proceed through gold movements.

In the early 2000s, the judgment on the international monetary system is mixed. On the one hand, it has cushioned major shocks such as the emerging-market crises of the late 1990s, the global-liquidity crisis following the collapse of LTCM in 1998, and the geopolitical risk arising from the terrorist attacks of 9/11. On the other hand, mounting current-account disequilibria in the US and China have fueled concern on the sustainability of the prevailing exchange-rate arrangements. The fact that the crisis eventually originated in the US housing market, not the foreign-exchange market, does not imply that global imbalances did not play a role—this is discussed further in chapter 8.

In this section, we first discuss the future of the international monetary system, then we outline the regional dimension of exchange-rate-regime choices.

a) Global imbalances

As noted earlier in this chapter (figure 5.6), the net foreign asset position of the US has deteriorated continuously in the 1990s and 2000s, with a current deficit that peaked above 6% of GDP in 2006. All other regions except Central and Eastern Europe have been in surplus, with more than half originating in Asia and oil-producing countries. Meanwhile, the composition of the external financing of the US has changed, with a relative diminution of private capital inflows and an increase in public financing, i.e., foreign central banks purchases of US securities. The People's Bank of China has played an important role in that respect. As a consequence, the international monetary conversation has gradually centered around the US and China. US trade deficits were mirrored by Chinese current-account surpluses, and the US financial account was partly financed by Chinese foreign-exchange-reserve accumulation.

The French economist and De Gaulle finance minister Jacques Rueff once described the Bretton Woods system as follows: "If I had an agreement with my tailor that whatever money I pay him returns to me the very same day as a loan," he said, "I would have no objection at all to ordering more suits from him." (Rueff, 1965). This was true under the Gold Standard of Bretton Woods until 1972, and again in the early twenty-first century, as a result of the fixed nominal exchange rate of the renminbi vis-à-vis the dollar.

A debate has developed in the early 2000s on the desirability and sustainability of this equilibrium. A first school of thought has argued that this *New Bretton Woods* or *Bretton Woods 2** (Dooley et al., 2004) was stable and reflected the structural growth patterns of the US and China.

Global imbalances would unwind naturally as a result of an accelerating supply in the US and stronger domestic demand in China.

Another line of thinking, more commonly found in international institutions such as the IMF has argued, in contrast, that the correction of global imbalances could be disorderly and lead to financial turmoil and a slowdown in world GDP. Policy action was therefore warranted to increase the saving rate in the US (by reducing the budget deficit and inciting households to save rather than spend out of their financial assets) and to decrease the saving rate in the rest of the world and particularly in China (by securing the domestic demand and extending the social safety net). This would reduce current-account imbalances and allow an orderly depreciation of the US dollar. In conjunction, China's exchange-rate regime would have to become more flexible so that the dollar can depreciate against the renminbi, not only against the euro and the yen.

Sustained global imbalances have resulted in an enormous accumulation of foreign-exchange reserves in some emerging market economies. At the beginning of 2008, official reserves of seven Asian countries (China, Japan, Taiwan, Korea, India, Singapore, and Hong Kong) totaled more than 3000 billion US dollars, i.e., more than 5% of world GDP. These countries had diversified their reserve portfolios away from US Treasury bonds and into other currencies and, increasingly, other asset classes such as bank deposits and government bonds. In the late 2000s, several countries including Russia and China had created so-called "sovereign-wealth funds" to invest their reserves in longer-term, more diversified portfolios, including stocks.[40]

The crisis that originated in the US in 2007 was mainly the result of misplaced incentives at a microeconomic level that led to excessive risk-taking by financial institutions. Whether it can also be partially ascribed to global imbalances which favored unbridled indebtedness in the US economy has been disputed. The G20 agenda for reform as outlined at the London Summit of April 2009 barely mentioned global imbalances and macroeconomic coordination (see chapter 8 for a discussion).

b) Regional monetary arrangements

Views on monetary regimes of emerging market economies have also evolved in the 2000s. As already discussed, the financial crises of the 1990s have made intermediary exchange-rate regimes outdated and induced a shift toward the extremes: Hard exchange rate pegs or floating rates. However, many emerging market countries have de facto pegged their exchange rate to the US dollar.

There are advantages and drawbacks to this situation. The absence of pre-announced nominal exchange-rate targets has made these countries less vulnerable to speculative attacks. However, pure dollar pegs have been

40. Some oil-exporting countries created such funds sometimes a long time ago, based on oil income.

increasingly inconsistent with the structure of trade and financial flows. The euro area weighs as much as the US in the foreign trade of many emerging countries, and regional trade integration has proceeded at an accelerated pace. Moreover, as already noted, most emerging countries have piled up current-account surpluses due to their high saving rates and, like China and Japan, they had to intervene by buying US dollars to prevent currency appreciation.

De facto pegs to the US dollar can be analyzed as a cooperative equilibrium in a game where each country is tempted to attract production away from its neighbors by depreciating its currency. If all countries follow this competitive devaluation policy, they will end up with an unchanged output and more inflation. It is an example of a prisoner's dilemma (see chapter 2). A first device to avoid such an outcome is policy coordination, which consists in deciding jointly on monetary policies, making it possible to internalize the impact of one's decisions on all other economies. However, coordination is costly: It entails information and negotiation costs, requires sanctions against noncooperative behavior, etc. If shocks are symmetric enough or economies are flexible enough, as explained above, another option is to create a currency area. Dollar pegs are a way to create a de facto currency area without incurring the cost of setting up new institutions.

Another option would be to introduce common pegs vis-à-vis a basket of key currencies, the structure of which would reflect trade and financial patterns between the region and the main economic areas. Ito and Ogawa (2002) propose such a system in the Asian case and prove that it would have better shock-absorbing properties. This could pave the way at a later stage to introduction of a single currency, Europe-style.[41] Such projects have been discussed in Asia (see Henning, 2009), in the Arab Gulf under the aegis of the Gulf Cooperation Council, and in Africa (box 5.14).

References

Asdrubali, P., B.E. Sørensen, and O. Yosha (1996), "Channels of interstate risk sharing: United States 1963–1990," *Quarterly Journal of Economics*, 111, pp. 1081–110.

Balassa, B. (1964), "The Purchasing Power Doctrine: A Reappraisal," *Journal of Political Economy*, 72, pp. 584–96.

Baldwin, R. (2006), "The Euro's Trade Effects," ECB working paper, no. 594.

Barisone, G., R.L. Driver, and S.W. Lewis (2006), "Are our FEERs justified?," *Journal of International Money and Finance*, 25, pp. 741–59.

Bayoumi, T., and B. Eichengreen (1994), "Shocking aspects of EMU," in Torres, F., and Giavazzi, F. (eds.), *Adjustment and Growth in the European Monetary Union*, Cambridge University Press.

Bénassy-Quéré, A., S. Béreau,and V. Mignon (2010), "On the Complementarity of Equilibrium Exchange-Rate Approaches," *Review of International Economics*, forthcoming.

41. See Cœuré (2004).

Bénassy-Quéré, A., and M. Coupet (2005), "On the Adequacy of Monetary Arrangements in Sub-Saharan Africa," *The World Economy*, 28, pp. 349–73.
Berg, A., and C. Pattillo (1999), "Are currency crises predictable?," *IMF Staff Papers*, 46, pp. 107–37.
Blanchard, O., and F. Giavazzi (2002), "Current Account Deficits in the Euro Area. The End of the Feldstein–Horioka Puzzle?," *Brookings Papers on Economic Activity*: 2.
Blanchard, O., and L. Katz (1992), "Regional Evolutions," *Brookings Papers on Economic Activity*, 1992–1, pp. 1–75.
Blanchard, O., and D. Quah (1989), "The Dynamic Effect of Aggregate Demand and Supply Disturbances," *American Economic Review*, 79, pp. 655–73.
Blanchard, O., F. Giavazzi, and F. Sa (2005), "International Investors, the US Current Account, and the Dollar," *Brookings Papers on Economic Activity*, 1, pp. 1–65.
Bordo, M., and T. Helbling (2003), "Have National Business Cycles Become More Synchronized?," *NBER Working Paper* No. 10130, Cambridge, MA: National Bureau of Economic Research.
Branson, W., H. Haltunen, and P. Masson (1977), "Exchange Rates in the Short Run," *European Economic Review*, 10, pp. 395–402.
Calvo, G., and C. Reinhart (2002), "Fear of Floating," *Quarterly Journal of Economics*, 117, pp. 379–408.
Cassel, G. (1921), *The World's Monetary Problems*. Constable and Co.
Chinn, M., and H. Ito (2008), "A new measure of financial openness," *Journal of Comparative Policy Analysis*, 10, pp. 309–22.
Clark, P., and R. MacDonald (1998), "Exchange Rate and Economic Fundamentals: A Methodological Comparison of BEERs and FEERs," *IMF Working Paper*, No. 98/00.
Clarke, S.V.O. (1967): *Central Bank Cooperation 1924–1931*, Federal Reserve Bank of New York.
Cœuré, B. (2004), "The narrow road to the single Asian currency," in De Brouwer, G., and Kawai, M. (eds.), *Exchange Rate Regimes in East Asia*, Routledge.
Corsetti, G. (1998), "Interpreting the Asian Financial Crisis: Open Issues in Theory and Policy," *Asian Development Review*,16, pp. 1–47.
Corsetti, G., and P. Pesenti (2001), "Welfare and Macroeconomic Interdependance," *Quarterly Journal of Economics*, 116, pp. 421–45.
Coudert, V., and C. Couharde (2005), "Real Equilibrium Exchange Rates in China," CEPII Working Paper 2005–01, January.
Dam, K. (1982), *The Rules of the Game: Reform and Evolution of the International Monetary System*, University of Chicago Press.
De Grauwe, P. (2000), "Exchange Rates in Search of Fundamentals: The Case of the Euro–Dollar Rate," *International Finance*, 3, pp. 329–56.
Dominguez, K., and J. Frankel (1993), *Does Foreign Exchange Intervention Work*? Washington DC: Institute for International Economics.
Dooley, M., D. Folkerts-Landau, and P. Garber (2004), "The Revived Bretton Woods System," *International Journal of Finance and Economics*, 9, pp. 307–13.
Eichengreen, B. (1992), *Golden Fetters: The Gold Standard and the Great Depression, 1919–39*, Oxford University Press.
Eichengreen, B., J. Tobin, and C. Wyplosz (1995), "Two Cases for Sand in the Wheels of International Finance, Editorial Note," *The Economic Journal*, 105, pp. 162–72.

Emerson, M., D. Gros, A. Italianer, J. Pisani-Ferry, and H. Reichenbach (1992), "One Market, One Money. An Evaluation of the Potential Benefits and Costs of Forming an Economic and Monetary Union," Oxford University Press.

Feldstein, M., and C. Horioka (1980), "Domestic Saving and International Capital Flows," *Economic Journal*, 90, pp. 314–29.

Flood, R., and P. Garber (1984), "Collapsing Exchange Rate Regimes: Some Linear Examples," *Journal of International Economics*, 17, pp. 1–13.

Flood, R., and A. Rose (1995), "Fixing the Exchange Rate: A Virtual Quest for Fundamentals," *Journal of Monetary Economics*, 36, pp. 3–37.

Fontagné, L., and M. Freudenberg (1999), "Endogenous Symmetry of Shocks in a Monetary Union," *Open Economies Review*, 10, pp. 263–87.

Fontagné, L., T. Mayer, and G. Ottaviano (2009), "Of Markets, Products and Prices: The Effects of the Euro on European Firms," EFIGE Report, Bruegel Blueprint, January.

Frankel, J. (1999), "No Single Currency Regime Is Right for All Countries or at All Times," *Essays in International Finance*, no. 215, Princeton University.

Frankel, J., and A. Rose (1996), "Currency Crashes in Emerging Markets: An Empirical Treatment," *Journal of International Economics*, 41, pp. 351–66.

Frankel, J., and A. Rose (1998), "The Endogeneity of the Optimum Currency Area Criteria," *The Economic Journal*, 108, pp. 1009–26.

Fratzscher, M. (2004), "Communication and exchange rate policy," ECB working paper, no. 363.

Frenkel, M., Ch. Pierdzioch, and G. Stadtmann (2001), "The Interventions of the European Central Bank: Effects, Effectiveness and Policy Implications," Working paper, Koblenz University.

Garber, P., and M. Taylor (1995), "Sand in the Wheels of Foreign Exchange Markets: A Sceptical Note,"*Economic Journal*, 105, pp. 173–80.

Garman, M., and S. Kohlhagen (1983), "Foreign Currency Option Values." *Journal of International Money and Finance*, 2, pp. 231–37.

Gaulier, G., A. Lahrèche-Révil, and I. Méjean (2008), "Exchange Rate Pass Through at the Product Level," *Canadian Journal of Economics*, 41, pp. 425–49.

Ghosh, A., A.-M. Gulde, J. Ostry, and H. Wolf (1997), "Does the Nominal Exchange Rate Regime Matter?," *NBER Working Paper*, no. 5874.

Giannone, D., and L. Reichlin (2006), "Trends and Cycles in the Euro Area: How Much Heterogeneity and Should We Worry About It?" *Working Paper Series*, No. 595, European Central Bank.

Goldstein, M. (2004), "China and the Renminbi Exchange Rate," in Bergsten, C.F., and J. Williamson (eds.), *Dollar Adjustment: How Far? Against What?*, Washington DC: Peterson Institute for International Economics.

Gourinchas, P.-O., and O. Jeanne (2006), "The Elusive Gains from International Financial Integration", *Review of Economic Studies*, 73, pp. 715–41.

Henning, R. (2009), "The Future of the Chiang Mai Initiative: An Asian Monetary Fund?," *Policy Brief*, Washington DC: Peterson Institute for International Economics, February.

HM Treasury (2003), "UK Membership of the Single Currency: An Assessment of the Five Economic Tests".

Ilzetski, E., C. Reinhart, and K. Rogoff (2008), "Exchange Rate Arrangements Entering the 21st Century: Which Anchor Will Hold?," available on Carmen Reinhart's Web site, http://terpconnect.umd.edu/~creinhar/.

Ilzkovitz, F., A. Dierx, V. Kovacs, and N. Sousa (2007), "Steps Towards a Deeper Economic Integration: the Internal Market in the 21st century, a Contribution to the Single Market Review," *European Economy*, no. 271, February.

International Monetary Fund (1997), "Capital Flows to Emerging Countries: a Historical Perspective," in *International Capital Markets: Developments, Prospects and Key Policy Issues*, September.

Ito, T., and E. Ogawa (2002), "On the Desirability of a Regional Basket Currency Arrangement," *Journal of the Japanese and International Economies*, 16, pp. 317–34.

Jeanne, O. (1996), "Les modèles de crises de change: un essai de synthèse en relation avec la crise du franc de 1992–1993," *Économie et Prévision*, no. 123–124.

Jeong, S., and J. Mazier (2003), "Exchange Rate Regimes and Equilibrium Exchange Rates in East Asia," *Revue Economique*, 54, pp. 1161–82.

Kalemli-Ozcan, S., B.E. Sørensen, and O. Yosha (2004), "Asymmetric Shocks and Risk Sharing in a Monetary Union: Updated Evidence and Policy Implications for Europe," *CEPR Discussion Paper*, no. 4463.

Kaminsky, G., and C. Reinhart (2003), "On Crises, Contagion and Confusion," *Journal of International Economics*, 51, pp. 145–68.

Kaminsky, G., S. Lizondo, and C. Reinhart (1998), "Leading Indicators of Currency Crises," *IMF Staff papers*, no. 45, pp. 1–48.

Kaplan, E., and D. Rodrik (2001), "Did the Malaysian Capital Controls Work?," *NBER Working Paper* no. 8142, Cambridge, MA: National Bureau of Economic Research.

Kenen, P. (1969), "The Optimum Currency Area: An Eclectic View," in Mundell, R., and Swoboda, A. (eds.), *Monetary Problems of the International Economy*, University of Chicago Press.

Kenen, P. (2000), *The International Economy*, Cambridge University Press, 4th edition.

Keynes, J.M. (1936), *The General Theory of Employment, Interest and Money*. Macmillan (reprinted 2007).

Krugman, P. (1979), "A Model of Balance of Payments Crises," *Journal of Money, Credit and Banking*, 11, pp. 311–24.

Krugman, P. (1987), "Pricing to Market when the Exchange Rate Changes," in Arndt, A., and Richardson, J. (eds.), *Real Financial Linkages Among Open Economies*, MIT Press, pp. 49–70.

Krugman, P. (1993), "Lessons of Massachussetts for EMU," in Torres, F., and Giavazzi, F. (eds.), *Adjustment and Growth in the European Monetgary Union*, Cambridge University Press, pp. 241–61.

Krugman, P. (1999), "Balance Sheets, the Transfer Problem and Financial Crises," in Isard, P., A. Razin, and A. Rose (eds.), *International Finance and Financial Crises: Essays in Honor of Robert Flood*, Kluwer Academic Publishers.

Lane, P. (2000), "The New Open Macroeconomy: A Survey," *Journal of International Economics*, 54, pp. 235–66.

Lane, P., and G. Milesi-Ferretti (2007), "The External Wealth of Nations Mark II: Revised and Extended Estimates of Foreign Assets and Liabilities, 1970–2004," *Journal of International Economics*, 73, pp. 223–50.

Lerner, A.P. (1944), *Economics of Control: Principle of Welfare Economics*, Macmillan.

Lucas, R. (1990), "Why Doesn't Capital Flow from Rich to Poor Countries?," *American Economic Review*, 80, pp. 92–96.

Lyons, R. (2001), *The Microstructure Approach to Exchange Rates*, MIT Press, p. 333.
Masson, P. (1999), "Contagion: Macroeconomic Models with Multiple Equilibria," *Journal of International Money and Finance*, 18, pp. 587–602.
Masson, P., and C. Pattillo (2001), "Monetary Union in West-Africa (ECOWAS): Is it Desirable and How Could it be Achieved?," IMF Occasional Paper, no. 204.
Mason, P., and C. Pattillo (2005), *The Monetary Geography of Africa*, Brookings Institution Press.
McKinnon, R. (1963), "Optimum Currency Areas," *American Economic Review,* 53, pp. 717–25.
McKinnon, R. (1993), "The Rules of the Game: International Money in Historical Perspective," *Journal of Economic Literature*, 31, pp. 1–44.
Mélitz, J., and F. Zumer (1999), "Interregional and International Risk-sharing and Lessons for EMU", Carnegie-Rochester Conference Series on Public Policy, 51, pp. 149–88.
Mohanty, M.S., and P. Turner (2006), "Foreign Exchange Reserve Accumulation in Emerging Markets: What are the Domestic Implications?," *BIS Quarterly Review*, September, pp. 39–52.
Mundell, R. (1961), "A Theory of Optimum Currency Areas," *American Economic Review*, 51, pp. 657–65.
Mundell, R. (1973), "Uncommon Arguments for Common Currencies," in Johnson, H., and Swoboda, A. (eds.), *The Economics of Common Currencies*, George Allen & Unwin Ltd, pp.114–32.
Mussa, M. (2000), "Factors Driving Global Economic Integration," Federal Reserve Bank of Kansas City Proceedings, pp. 9–55.
Mussa, M., A. Swoboda, J. Zettelmeyer, and O. Jeanne (2000), "Moderating Fluctuations in Capital Flows to Emerging Economies," in Kenen, P., and Swoboda, A. (eds.), *Reforming the International Monetary and Financial System*, Washington: International Monetary Fund.
Obstfeld, M. (1994), "The Logic of Currency Crises," Banque de France, *Cahiers économiques et monétaires*, 43, pp. 189–213, and NBER Working Paper, 4640.
Obstfeld, M., and K. Rogoff (1995), "Exchange Rate Dynamics Redux," *Journal of Political Economy*, 103, pp. 624–60.
Obstfeld, M., and K. Rogoff (1995), "The Intertemporal Approach to the Current Account," in Grossman, G., and K. Rogoff, *Handbook of International Economics*, volume 3, chapter 34, pp. 1731–99, Elsevier.
Obstfeld, M., and K. Rogoff (1999), *Foundations of International Macroeconomics*, MIT Press.
Obstfeld, M., and K. Rogoff (2007), "The Unsustainable US Current Account Position Revisited", in R. Clarida (ed.), *G7 Current Account Imbalances: Sustainability and Adjustment*, University of Chicago Press.
Obstfeld, M., and A. Taylor (2004), *Global Capital Markets: Integration, Crisis, and Growth*, Cambridge University Press.
OECD (2002), *Economic Outlook* 71, chapter 6, Paris: OECD.
Poole, W. (1970), "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic Macro Model," *Quarterly Journal of Economics*, 84, pp. 197–216.
Robinson, J. (1947), *Essays in the Theory of Employment*, Part III, "The Foreign Exchanges," Blackwell.
Rogoff, K. (1996), "The Purchasing Power Parity Puzzle," *Journal of Economic Literature*, 34, pp. 647–68.

Rose, A. (2000), "One Money, One Market: The Effects of the Common Currency on Trade, *Economic Policy*, 30, pp. 7–47.
Rubin, R., and J. Weiseberg (2003), *In an Uncertain World: Tough Choices from Wall Street to Washington*, Random House Trade Paperback Edition.
Rueff, J. (1965), "Return to Gold—Argument with Jacques Rueff," *The Economist*, 13 February.
Sachs, J., A. Tornell, and A. Velasco (1996), "Financial Crises in Emerging Markets: the Lessons from 1995," *Brookings Papers on Economic Activity*: 1, Brookings Institution, pp. 146–215.
Salant, S., and D. Henderson (1978), "Market Anticipation of Government Policies and the Price of Gold", *Journal of Political Economy*, 86, pp. 627–48.
Samuelson, P. (1964), "Theoretical Notes on Trade Problems," *The Review of Economics and Statistics*, 46, pp. 145–54.
Sarno, L., and M. Taylor (2001), "Official Intervention in the Foreign Exchange Market: Is It Effective and, If So, How Does It Work?," *Journal of Economic Literature*, 39, pp. 839–68.
Sørensen, B., and O. Yosha (1998), "International Risk-Sharing and European Monetary Unification," *Journal of International Economics*, 45, pp. 211–38.
Stein, J. (1994), "The Natural Real Exchange Rate of the US Dollar and Determinants of Capital Flows," in J. Williamson (ed.), *Estimating Equilibrium Exchange Rates*, Washington DC: Peterson Institute for International Economics.
Summers, L. (1988), "Tax-policy and International Competitiveness," in Frenkel, J. (ed.), *International Aspects of Fiscal Policy*, University of Chicago Press, pp. 349–75.
Taylor, A. (1996), "International Capital Mobility in History: The Saving-Investment Relationship," NBER Working Paper, 5743.
Tobin, J. (1958), "Liquidity Preference as Behaviour Towards Risk," *Review of Economic Studies*, 25, pp. 65–86.
Triffin, R. (1960), *Gold and Dollar Crisis*, Yale University Press.
Williamson, J. (1983), *The Exchange Rate System*, Washington DC: Institute for International Economics.

6

Growth Policies

6.1 Issues
- 6.1.1 Five "stylized facts"
- 6.1.2 Catching up
- 6.1.3 The origin of productivity differentials

6.2 Theories
- 6.2.1 Growth through capital accumulation
- 6.2.2 External effects, innovation, and growth
- 6.2.3 Beyond the production function

6.3 Policies
- 6.3.1 A roadmap
- 6.3.2 Improving institutions
- 6.3.3 Investing in education, innovation, and infrastructures
- 6.3.4 Increasing labor supply
- 6.3.5 Making labor and product markets work better
- 6.3.6 Developing and regulating financial markets
- 6.3.7 Countering the effects of distance and history
- 6.3.8 Choosing priorities

References

In the previous chapters, we discussed the role of economic policies in managing aggregate demand. Yet, while such demand policies may be effective in the short term to dampen cyclical fluctuations, they are unable to foster lasting growth. Long-term production is essentially determined by potential supply, whose rate of growth conditions the increase in wealth and welfare. People benefit from it directly through higher incomes or indirectly through wider access to public goods such as health, education, safety, and infrastructures. They also suffer from some of its consequences, such as damage to the environment.

However, economic growth is by no means a foregone conclusion. In 1913, Argentina's gross domestic product per person was 70% more than Spain's.[1]

1. The relevance of gross domestic product as a measure of a country's standard of living will be discussed in the first section of this chapter.

In the aftermath of World War II, that of Ghana exceeded that of South Korea by almost 50%. In 1970, Italy's GDP per person was more than 60% higher than Ireland's. Yet, at the turn of the millennium, Spain's GDP per person was 50% higher than Argentina's, Korea's was 10 times higher than Ghana's, and Ireland's just exceeded Italy's.[2] Prosperity and underdevelopment thus result from persistent growth divergences over decades. For example, a growth differential of a single percentage point per year (as compared to a 2% benchmark) cumulated over 50 years results in a 63% income gap; for a differential of two percentage points, the gap after 50 years reaches 164%; and for three percentage points, it is 326%.[3] On such arithmetic, based on simple assumptions regarding productivity and demographic trends, Fogel (2007) expects the three largest economies in the world in 2040 to be China (expected to represent 40% of the world economy in purchasing-power parity terms), the US (14%) and India (12%).

The divergence of growth rates over time and across space is a well-documented fact. Trade and capital flows have not equalized growth rates. How to promote economic growth stands as one of the main economic policy issues and probably the most important one from a long-term perspective. It is also one of the most difficult to address: The quest for the determinants of growth is similar to opening successively, and with increasing difficulty, a set of Russian dolls. Opening each doll highlights a part of reality, but it is accompanied by the realization that deeper insights are necessary but still out of reach. Yet, it is on such insights that economic policy should be based in order to stimulate sustainable growth.

A first step toward understanding the growth process consists in documenting the trajectories of various countries' economies over time. A number of stylized facts emerge from this exercise. A second step consists in uncovering the determinants of economic growth by introducing *production factors**, namely labor and capital, and by calling on what is known as "growth accounting," which identifies their contribution to economic growth. This approach, however, remains descriptive. To proceed further, it is necessary to turn to growth models. So-called "neoclassical" growth models explain human and physical capital accumulation and its impact on income per person. They make growth depend on the behavior of savings and on investment in education. These models are closer to reality—they help understand why growth rates in Europe and Japan slowed down at the end of the twentieth century when income per person in those economies got closer to the US level—and help address economic policy choices. However, they remain frustrating in that a significant part of the determinants of long-term growth comes out as an unexplained residual. To understand what is behind that residual, later models have focused on endogenous growth mechanisms that explain how growth can persist over time and regenerate,

2. These comparisons are drawn from Angus Maddison's work (2001) on growth in the long term.
3. The exact figures, of course, depend on the rate of growth used as a benchmark.

and that help understand why growth performance differs across countries of similar development levels. These models, whose aim is to explain why some countries make better use of their production factors than others, need strong microeconomic underpinnings. In the words of the Spence Commission, which produced an influential report on long-term growth in developing countries, "the growth of GDP may be measured up in the macroeconomic treetops, but all the action is in the microeconomic undergrowth, where new limbs sprout, and dead wood is cleared away." (Commission on Growth and Development, 2008, pp. 43–44). In turn, these models point to deeper determinants such as economic institutions and their adequacy for a given level of development. The last stage of the analysis therefore consists in determining in what context (in terms of education and research systems, of intellectual property protection, of competition, of corporate finance, of taxation, etc. ...) innovation and endogenous growth are most likely to flourish.[4]

Moving to a deeper level of analysis does not deprive the previous one of relevance, and the various approaches we have outlined are to a large extent complementary. To start with, as developed below, basic models go a long way toward explaining growth differences: For example, the largest part of the stellar East Asian growth performance is accounted for by high saving and investment rates. Second, as economists move away from the well-charted approaches based on measuring growth and its components toward understanding its deeper determinants, their knowledge and recommendations become less assertive. Informed by past errors, they realize the limits of their science.[5] However, at the same time, the challenges of economic policymaking underline how essential such investigations are. To find out how to jumpstart European growth, for example, it is not enough to observe that it has been diverging from that of the US, nor even that investment has slowed down or that innovation has lagged behind: What matters is to determine whether European countries should as a priority devote additional resources to education and research, whether they should enforce tighter competition in product markets, or whether they should embark on tax reforms.

4. Among the explanatory factors of growth, some are beyond the economists' realm, because they are exogenous to the operation of the economy. For example, landlocked countries face significant challenges—which does not mean they cannot develop, depending on the natural resources they have and how well they and their neighbors perform (just compare Switzerland and Rwanda).

5. In the 1950s, there was widespread pessimism with respect to Asian development prospects. In the 1960s, the idea that the Soviet Union was on its way to catching up with the US was commonly accepted. In the 1970s, European countries seemed to have definitely entered a high growth path. In the 1980s, Japan was regarded as a model while the US economy seemed to be plagued by deindustrialization and declining productivity. The 2007–09 financial and economic crisis may lead in retrospect to a more critical diagnosis of economic policies in the 1990s and 2000s and of the underlying health of the US economy in these decades of rapid growth.

Growth economics therefore relies on a combination of mechanics and alchemy. The former are necessary to isolate and quantify the proximate determinants of output development along a growth trajectory. The latter is called for to understand what makes countries take off and move from one growth trajectory to a higher one. Beyond the obvious—the recommendation to fix all the major deficiencies that hinder labor market participation, the acquisition of knowledge, capital accumulation, innovation, and productivity improvements—there is therefore no such thing as a growth recipe. This makes the search for successful growth strategies especially arduous. To borrow from the title of a book by a famous development scholar (Easterly, 2001), the quest for growth is bound to remain elusive.

6.1 Issues

6.1.1 Five "stylized facts"

Five stylized facts emerge from observation:

- By historical standards, fast growth in income per person is a recent phenomenon;
- Along a growth path, income per person and productivity exhibit significant medium-term turning points that are not necessarily synchronous across countries at similar development levels;
- Convergence at the top is neither general nor unattainable. In the last decades, the income per person in some formerly underdeveloped countries, such as East Asian countries, has caught up with that of the most advanced ones, but other countries, including most sub-Saharan African countries, have further diverged;
- Largely as a consequence of growth developments, income inequalities among world citizens increased strongly during the nineteenth and the first half of the twentieth centuries. They have stabilized since the 1990s, essentially through the rapid increase in wealth of part of the Chinese and Indian populations;
- Growth patterns differ over time and they can at times increase inequality within countries.

Before presenting each of these stylized facts, a short methodological discussion is required.

a) How can growth and development be measured?

"A rising tide lifts all boats," John F. Kennedy once famously said.[6] But all boats are not lifted equally. Aggregate measures of the level of development require aggregating the standards of living of many individuals and such

6. Remarks in Heber Springs, Arkansas, at the Dedication of Greers Ferry Dam, 3 October 1963.

aggregation involves ethical choices: The utilitarian, or "Benthamian" (cf. chapter 1) observer will assess development through the well-being of the community as a whole and therefore through the evolution of average income, while the "Rawlsian" observer will focus on the poorest individual and will therefore be concerned with absolute poverty reduction.[7] The Rawlsian criterion emphasizes social justice and it is often supported by the free-marketers, who are willing to accept a deepening of inequality as long as the poorest also benefit. Macroeconomists typically adopt a utilitarian approach, since they are chiefly concerned with the progression of income per person and often ignore income distribution. Development economists often blend the two approaches by looking both at average income and at some measure of absolute poverty, such as the proportion of the population living on less than one dollar per day.

*GDP per person**, or *GDP per capita**, corresponds to the average value added per person created within a given constituency and is therefore relevant to measuring the average standard of living, however equitable its distribution may be.[8]

Comparing countries' incomes at various points in time requires a number of technical corrections. Across time, GDP needs to be measured *in constant prices**, that is, to be adjusted for the evolution of prices. Across countries, adjustments are also needed to account for variations in the exchange rate. The common practice is to use exchange rates adjusted for price differences: *Purchasing-power parity exchange rates* or PPP exchange rates* are the nominal exchange rates that would equalize prices across countries.[9] These can be computed but they are fraught with uncertainty which significantly affects comparisons.[10]

Furthermore, per-capita GDP is beset by a number of shortcomings, further discussed in the report by the Commission on the Measurement of Economic Performance and Social Progress (Stiglitz et al., 2009, cf. chapter 1). First, GDP

7. An observer who would be concerned with the distribution of welfare would stand between the two. On these issues, see Amartya Sen's Nobel lecture (Sen, 1999).

8. We do not discuss here the difference between per-capita GDP—which measures the average output per person produced by the residents of a given territory—and GNP per person—which measures the *residents*' average income. Both measures can differ appreciably when residents own external assets (or conversely when they are indebted to nonresidents), when they receive private transfers from foreign countries (in particular from emigrants working abroad), or when they benefit from international development assistance. Our purpose here is not to analyze these differences and we shall therefore use income per person and per-capita GDP interchangeably.

9. Purchasing power parity is defined in chapter 5.

10. PPP rates are published by the World Bank, which coordinates an international comparison program (ICP), based on price surveys conducted at three-to-five-year intervals, as well as estimations for some countries. One of the notable innovations in the 2006 ICP program was the full participation of China, which provided price surveys yielding a more accurate estimate of the country's PPP exchange-rate-based GDP. As a consequence of prices being higher than previously thought, China's real GDP was revised downward by about 40%, which in turn affected the measurement of world growth by half a percentage point over the 2005–08 period. This episode is a strong reminder of the fragility of international comparisons.

sums up individual expenditures and does not take into account positive or negative externalities: For example, the value added that is generated by polluting industries adds to GDP, but the damage they cause to the environment is ignored—on the contrary, expenditures made necessary to correct this damage contribute to GDP. Individual welfare also depends on life expectancy, on access to public services, on the length and quality of leisure, etc. As indicated in chapter 1, GDP does not measure welfare accurately. In the words of the economist and philosopher Amartya Sen, who was awarded the Nobel Prize for his research on social justice:

> Rather than concentrating only on some solitary and traditional measure of economic progress (such as the gross national product per head), 'human development' accounting involves a systematic examination of a wealth of information about how human beings in each society live. . . . Human lives are battered and diminished in all kinds of different ways, and the first task, seen in this perspective, is to acknowledge that deprivations of very different kinds have to be accommodated within a general overarching framework. The framework must be cogent and coherent, but must not try to overlook the pluralities that are crucially involved (in the diverse nature of deprivations) in a misguided search for some one measure of success and failure, some single clue to all the other disparate concerns.
>
> Amartya Sen (2000, p. 18)

The United Nations Development Program (UNDP) has introduced a human development indicator to provide a more comprehensive measure of well-being (HDI, described in chapter 1).

Second, conventional GDP does not take into account the depletion of natural resources. The concept of *sustainable development** aims at correcting this shortcoming through introducing intertemporal concerns and taking into account the way in which current patterns of production and of consumption will affect those of tomorrow. While various efforts have been made to develop integrated so-called environmental-economic or "green" national accounts,[11] however, no single headline indicator is yet available that adequately captures this intertemporal dimension within national accounts.

Third, per-capita GDP is not relevant for studying the efficiency of production, because a number of national residents do not work. Productive efficiency is better captured by *labor productivity** (cf. box 6.1).

11. For example, the United Nations Statistical Commission established in 2005 a Committee of Experts on Environmental-Economic Accounting (UNCEEA) in order to mainstream environmental-economic accounting and establish a system of integrated environmental and economic accounting as an international standard. See also Hamilton (2006) for estimates of produced, natural, and intangible capital as well as "genuine" savings rates that take environmental degradation into account.

Box 6.1 From Per-Capita GDP to Labor Productivity

Per-capita GDP is the ratio of the value added Y created during a given year in a given country (the Gross Domestic Product) to the country's total population *Pop*. It depends on productivity but also on participation, employment and hours worked:

- A fraction $(1 - x)$ of the population (children, students, retirees, but also adults excluded from the labor market such as invalids, housewives, etc) does not participate in the labor market. x is called the *participation rate**.
- The *labor force** is therefore $L = xPop$. Within it, however, a fraction u is unemployed and *employment** is $N = (1 - u)xPop$.
- Lastly, each employee works on average d hours. The total number of hours worked is therefore $H = d(1 - u)xPop$.

Finally, labor productivity is:

$$\frac{Y}{H} = \frac{1}{1-u}\frac{1}{d}\frac{1}{x}\left(\frac{Y}{Pop}\right) \qquad \text{(B6.1.1)}$$

This is sometimes adjusted for labor quality to account for divergences in skills. Table B6.1.1 decomposes the gap between the number of hours worked in the euro area and in the US in 2008. While the European population was 6% larger, the quantity of labor supplied was 17% lower. This divergence stemmed from differences in the average number of hours worked d (1792 hours a year in the US, 1574 in the euro area), in the participation rates x and, to a lesser extent, in the unemployment rates u.

Table B6.1.1
Number of hours worked in 2008 in the Euro area and in the US

	Variable	US	Euro area	Euro area versus US
Total population in millions	*Pop*	304	322	+6%
Ratio 15–64-year-old/total population		67%	67%	—
Participation rate of the 15–64-year-olds	x	75%	73%	–3%
Employment rate	$1 - u$	94%	92%	–2%
Average number of hours worked	d	1792	1574[a]	–12%
Total number of hours worked (billion)	H	259.8	226.7	–13%

Notes: Civilian employment only. [a]Weighted average of the four largest countries.

Source: OECD, *Labor Force Statistics* 2009.

Consequently, in the comparison between the euro area and the US, there is a factor of $(1792/1574) \times (0.94/0.92) \times (0.75/0.73) \times (0.67/0.67) = 1.20$ between the gap in per-capita GDP and the gap in GDP per hour worked.

This difference is particularly relevant when comparing economic performance between Europe and the US. Table B6.1.1 indicates that in 2008, while the population of the euro area was 6% higher than the US population, the total number of hours worked in the euro area was 13% lower. This 19% gap chiefly came from the number of hours worked by employees and from the lower proportion of persons in the labor force. Consequently, the comparative judgment differs depending on whether GDP per person or labor productivity is considered. The euro area's GDP per person is 23% lower than that of the US (see figure 6.2 below) but labor productivity is only 3% lower. Olivier Blanchard (2004) has argued that the gap in GDP per person can therefore not be ascribed to any difference in economic performance, but rather points to a European "preference for leisure." The adequacy of this diagnosis has, however, been questioned. In particular, productivity in the euro area may be over-estimated because a large proportion of low-skill workers whose potential productivity is below average are excluded from the labor force.[12]

This chapter does not address measurement issues any further, but the theoretical tools that are introduced below are compatible with potential extensions of the concepts of growth and development.

b) Trends and turning points

Angus Maddison, a renowned scholar of the quantitative history of growth, has built an indicator of world GDP per person since year 1 CE and has projected it to 2030 (Maddison, 2007). The same indicator has been extrapolated to one million years BCE by assuming a stable link between population and standard of living before the industrial revolution (Kremer, 1993).[13]

Figure 6.1 shows the world GDP per person (in 1990 purchasing-power-parity dollars) since the start of the first millenium. Four major periods can be distinguished. From prehistory through the Middle Ages, yearly income remains at around $450 per person (in fact, it declines throughout the first millenium). It then increases to about $600 between 1400 and 1800. The true "take-off" comes with the industrial revolution in the nineteenth century and GDP per person exceeds $1500 on the eve of the World War I. By 2003 it reaches $6500, having multiplied by more than five over the course of the century. Maddison expects it to reach $11700 by 2030.

World income per person therefore experienced a long period of stability followed by a take-off. Its growth appears as a recent phenomenon: This is the *first stylized fact.*

Turning points[14] can be related to changes in the world economic system and especially to breakthroughs that have been conducive to productivity

12. See Cette (2004, p. 24, Table 4).

13. Up to1800, a 1%-a-year acceleration of population growth is associated with a $1165 rise in per-capita GDP.

14. One of the most important of these turning points was the rise of Europe. The US physiologist Jared Diamond (1997) has proposed a pioneering explanation that opened a lively debate.

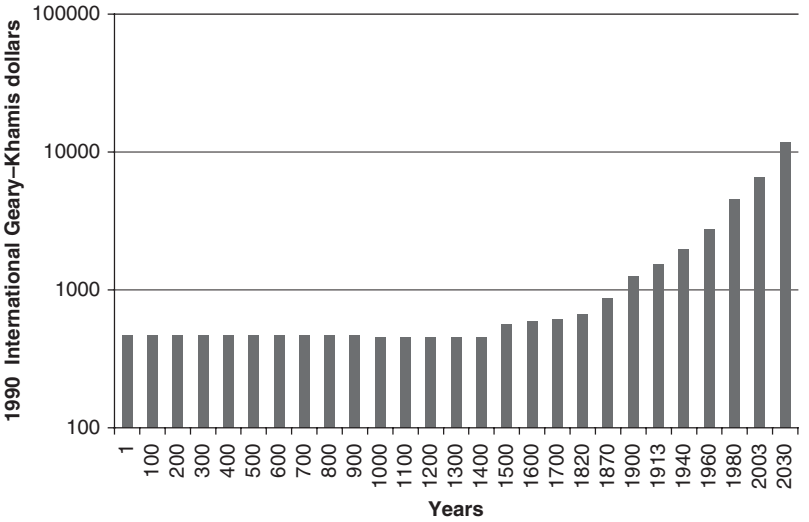


Figure 6.1 Long-term evolution of world GDP per person.
Source: Data from Maddison (2007), http://www.ggdc.net/maddison/.

Note: GDP per person in 1990, purchasing-power parity, international (Geary–Khamis) dollars.

and international trade: Improvements in agricultural productivity; the "discovery" of America in the fifteenth and sixteenth centuries; the emergence and expansion of the so-called "European world-economy";[15] major technological innovations such as the steam engine and the railways in the nineteenth century; electricity in the twentieth century; and urbanization on a large scale. However, technology and trade alone hardly account for the recent dramatic increase in world GDP. They cannot explain why the standard of living did not increase until the Middle Ages, in spite of a string of technological innovations (fire, the wheel, metals, and later navigation). Understanding these turning points involves the study of history as much as economics.[16]

What happens in the second half of the twentieth century? In the 1950s and 1960s, Europe and Japan rapidly caught up with the US economy (figure 6.2a). In the 1970s and 1980s, the three major economies slowed down in the

He has assigned the European successes to the development of agriculture and of livestock-farming, themselves due to the local abundance of seeds and the availability of animals which could be domesticated, allowing the growth of productivity and the greater concentration of people. Europe's East–West geography facilitated migration within a constant climatic environment, and therefore innovation and technological diffusion. Also, proximity with domesticated animals could have allowed the immunization of local people against microbial germs. When physical contact between the Europeans and indigenous populations in other continents took place on the occasion of major explorations, the latter suffered from pandemics while the former remained immune from them.

15. On the world-economy concept, see Braudel (1981–84, vol. 3, ch. 1) and Wallerstein (1979).
16. For a synthetic presentation, see Braudel (1985).

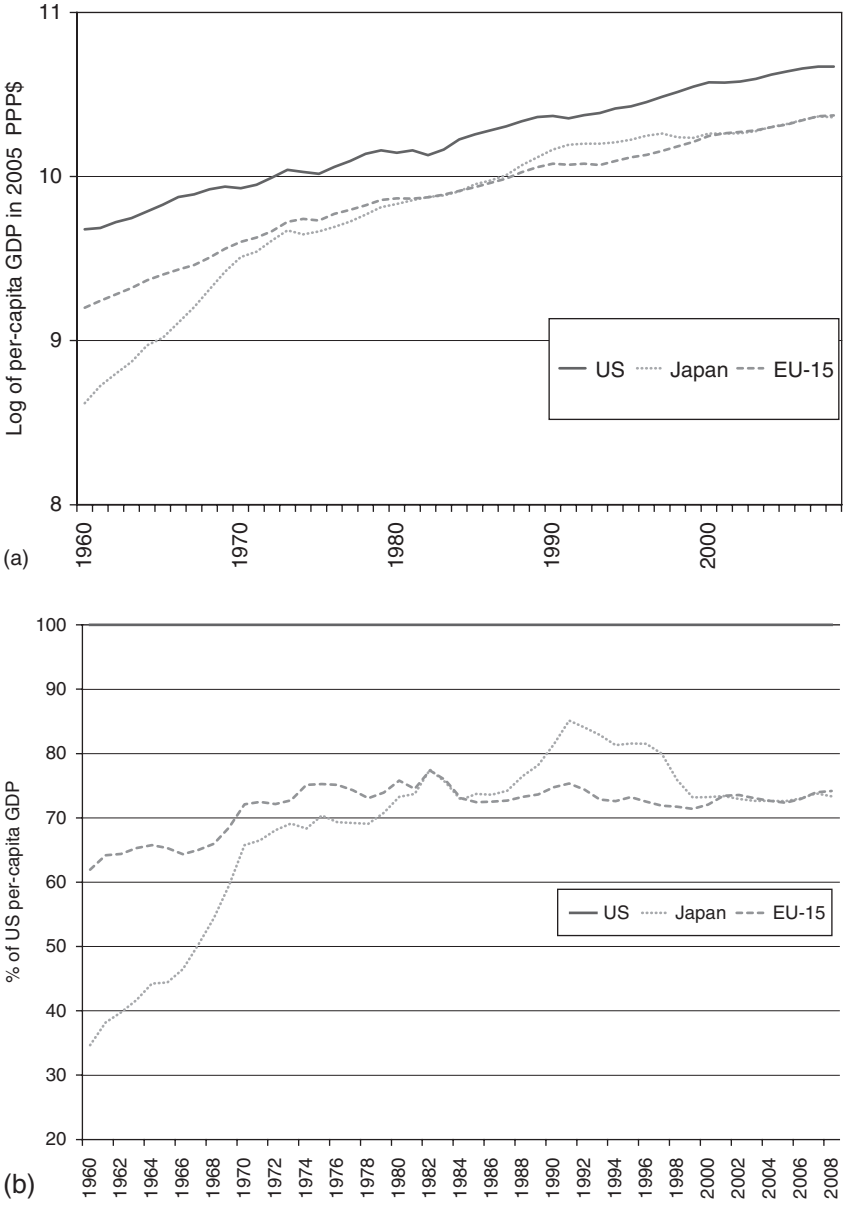


Figure 6.2 Real GDP per person in the EU, the US, and Japan, in 2005 purchasing power parity USD. a) Per-capita GDP levels in logarithms, b) in percent of US per-capita GDP.
Source: OECD database, April 2010.

Note: There is a break in the European series in 1991 because of German unification.

aftermath of the oil shocks but the US economy accelerated again in the 1990s, with the result that catching up by Europe and Japan came to an end and even started to reverse. Convergence stopped at 80% of US income per person. This *relative* evolution is shown on figure 6.2b which shows the same data as in figure 6.2a, but as a percentage of the US level. It would appear even more clearly if one considered only Germany, France, or Italy.

GDP per person and productivity can thus experience significant synchronous and asynchronous inflections. Especially Europe and Japan, which had been catching up with the US standard of living since the Second World War, fell behind after the 1980s. This is our *second stylized fact.*

c) Convergence and inequality

Who benefits from the increase of world income and wealth? This is an important question from the standpoint both of economic dynamics and of political economy (because individuals vote on economic policies). Economic growth and income inequality interact in two natural ways: Inequalities affect growth; and growth influences inequality.

Both channels can be analyzed from a purely positive point of view, asking whether inequality is good or bad for economic growth and whether growth leads to increasing or reducing inequality. Both are fiercely debated among economists because several effects interplay, the combined result of which is uncertain, as discussed in the second section of this chapter. However, inequality inevitably also involves normative judgments and the associated diagnosis will depend on the implicit or explicit concept of inequality one focuses on—for example, income dispersion or a Rawlsian emphasis on the lowest incomes. Finally, the conclusion will be different whether one considers inequality across countries, or inequality between individuals within countries, or global inequality among world citizens irrespective of the country they belong to (which depends on both inequality across and within countries).

The concern about income inequality rightly mobilizes attention. However, in the long term, improvements in livelihoods brought by technical progress and economic growth play a decisive role in the evolution of income. Robert Lucas draws on this observation to caution against excessive confidence in policies of sheer income redistribution:

> In this very minute, a child is being born to an American family and another child, equally valued by God, is being born to a family in India. The resources of all kinds that will be at the disposal of this new American will be on the order of 15 times the resources available to his Indian brother. This seems to us a terrible wrong, justifying direct corrective action, and perhaps some actions of this kind can and should be taken. But of the vast increase in the well-being of hundreds of millions of people that has occurred in the 200-year course of the industrial revolution to date, virtually none of it can be attributed

> to the direct redistribution of resources from rich to poor. The potential for improving the lives of poor people by finding different ways of distributing current production is *nothing* compared to the apparently limitless potential of increasing production.
>
> Robert Lucas (2004)

From an international perspective, the argument implies that the improvement in poor people's incomes relies on their country's economic growth rather than on the *direct* effects of official development assistance. However, it would be wrong to conclude that aid does not contribute to economic growth and development.

In what follows, we first discuss income inequality between countries, then inequality within countries and among world citizens.

As noted above, growth in GDP per person has been exponential, which implies that small growth differentials between countries have resulted over the long term in very significant divergences in income levels. Figure 6.3 illustrates the effects of these cumulated growth divergences for a sample of countries. On the whole, a positive correlation can be observed between initial and final GDP per person, with a less-than-unitary slope of the regression line indicating that poorer countries have on average grown faster. However, some countries were poor in 1870 but rich in 2000, or vice versa. For example, Uruguay was three times richer than Japan in 1870 but almost three times poorer in 2000. This inversion was the result of differential growth rates cumulated over a long period.

A number of countries have been completely left out of the growth process. According to the United Nations Development Program (2007), in 2005 the dispersion of GDP per person expressed in purchasing-power-parity dollars ranged from 1 to 90 between Malawi ($667 per person and per year) and Luxembourg ($60228), respectively. The five richest countries were four European countries and a country of European immigration: Luxembourg, the US, Norway, Ireland, and Iceland. The five poorest countries were all African (Malawi, Burundi, Democratic Republic of Congo, Tanzania, and Niger). Not all African countries, fortunately, suffer a stagnation in *absolute* terms, but their exclusion in *relative* terms is confirmed by the comparison of their per-capita GDP with the US level in the second half of the twentieth century: Whereas some convergence occurred in Europe (figure 6.2) and in Asia (figure 6.4a), many African countries have progressed in absolute terms but not in relative terms (figure 6.4b) and the promising resumption of economic growth in sub-Saharan Africa from 1995 on, halted by the 2007–09 crisis, was still insufficient to reverse that trend.[17]

17. However, Sala-i-Martin and Pinkovskiy (2010) challenge the presumption that Africa has not reduced poverty and argue instead that African poverty has been falling rapidly for all classes of countries.

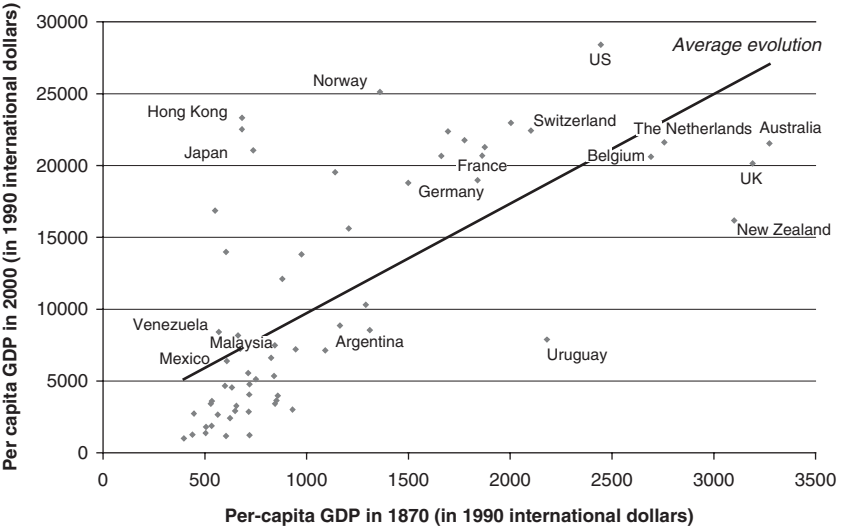


Figure 6.3 GDP per person in 1870 and 2000.
Source: Data from Maddison (2007), http://www.ggdc.net/maddison/.

Note: GDP per person in 1990, purchasing-power parity, international (Geary–Khamis) dollars.

However, as Dani Rodrik (2005) observes, developing countries experienced between 1960 and 2000 a historically strong rate of growth of their income per person, at a rate of 2.3% a year (against 1.3% for England throughout the period of its economic supremacy between 1820 and 1870) but the income per person grew even faster over the same period in the rich countries, at a rate of 2.7% a year. Figure 6.4 illustrates these very different outcomes in terms of catching up.

Convergence of GDP-per-person levels has taken place within certain groups of countries but is by no means a general phenomenon. On the contrary, *some countries have kept out of the dynamics of convergence and even further diverged*: This is the *third stylized fact.*

Let us now turn to individual incomes. In 1955, Simon Kuznets suggested that there was an inverted U-shaped relationship between the level of development and within-country income inequality: Inequality would be low in poor countries (like African countries) and in rich countries (like Europe), but high in those in between (like Latin American countries). As a consequence, development came together with a temporary rise in inequality.

The *Kuznets curve** was influential in shaping views on the trade-offs implied by development but it is empirically disputed (Deininger and Squire, 1996) and rests on unclear theoretical foundations. Kuznets explained it by the reallocations of labor during the industrial transition phase. In an agrarian economy, income inequalities are weak; in the first phase of the transition, polarization between the agricultural sector and the manufacturing/urban

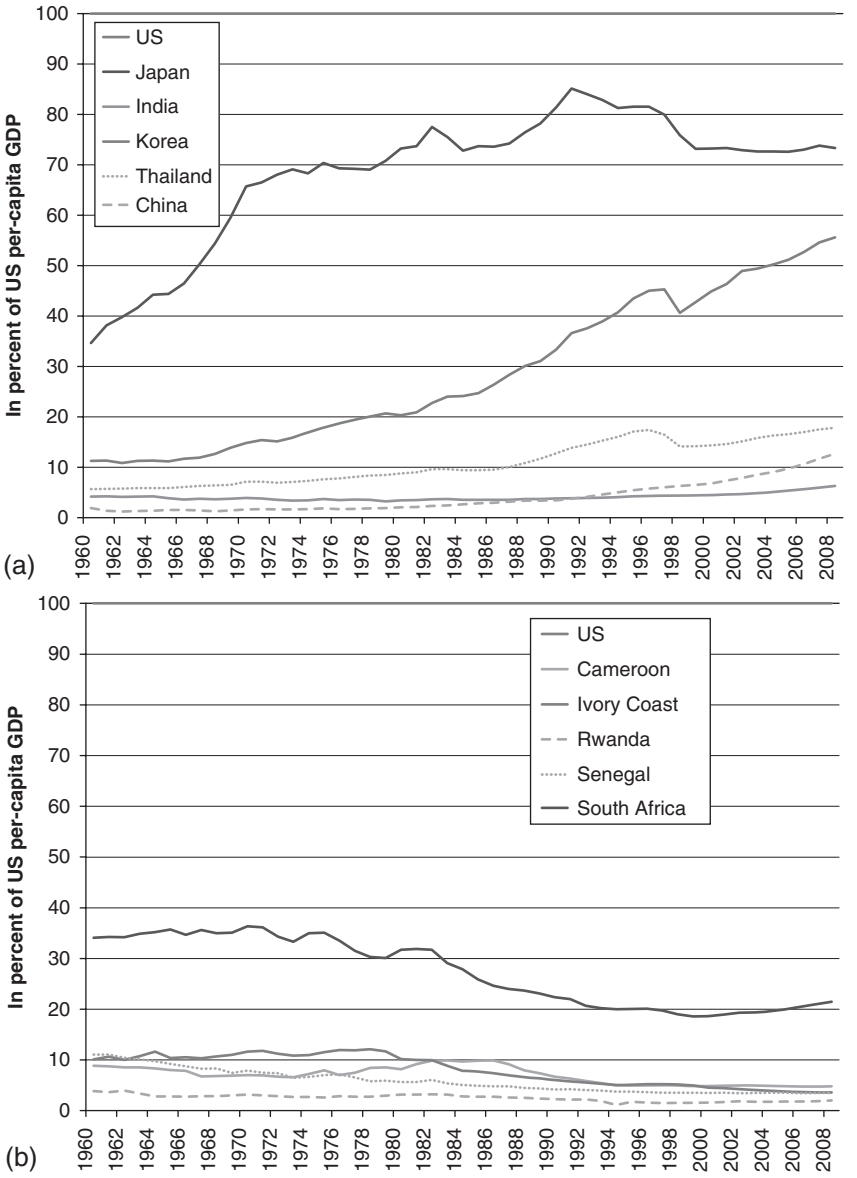


Figure 6.4 Per-capita GDP relative to the US level, in purchasing power parity (USD of 2005). a) Asia, b) Africa.
Source: IMF, *World Development Indicators* database, April 2010 .

Note: GDP per person in 1990, purchasing-power parity, international (Geary–Khamis) dollars.

sector increases inequalities; and as the agricultural sector shrinks over time, inequalities are reduced. The very sharp increase in world inequality during the nineteenth and twentieth centuries (box 6.2) seems to validate this explanation. Modern thinking emphasizes the income effects of within-sector technological innovation, rather than inter-sectoral dimensions, as innovation creates temporary but unequally distributed rents (Galor and Tsiddon, 1997). The initial rise in inequality comes as a corollary to the *creative destruction* process that will be discussed in the second section of this chapter.

Box 6.2 Global Inequality

François Bourguignon and Christian Morrison (2002) have studied the world distribution of individual incomes over a very long period, between 1820 and 1992. Further work by Branco Milanovic (2005) provides refinement and an update of this research.

Global inequality increased considerably during the last two centuries. The *Lorenz curve* describing the cumulated distribution of income (cf. chapter 1) has gradually further departed from the 45° line (figure B6.2.1), which indicates a concentration of income among the richest individuals. The *Gini index* of world income distribution[a] increased from 0.5 in 1820, to 0.64 in 1950, and 0.66 in 1992. It has remained stable since then (Milanovic, 2005).

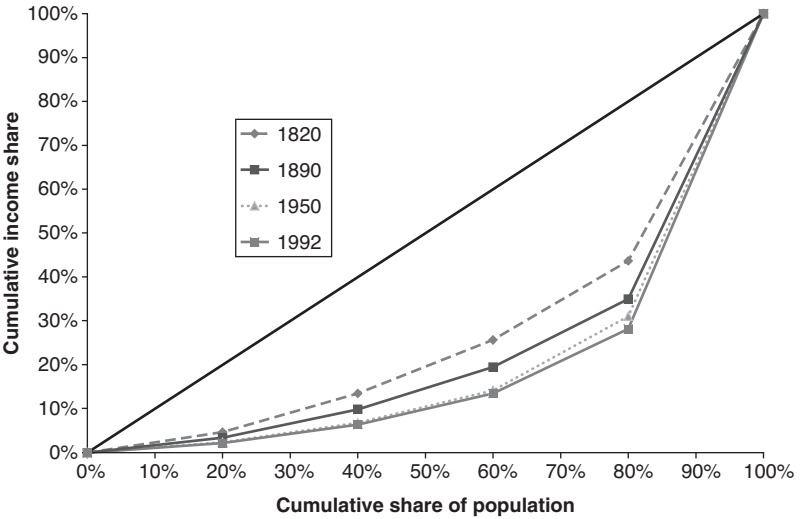


Figure B6.2.1 Lorenz curve of world income.
Source: Bourguignon and Morrison (2002), table 1.

Reading: The poorest half of the world population received approximately 20% of world income in 1820, but only 10% in 1992.

The increase in global inequality since the early nineteenth century is attributable to the rise of inequality between countries. In fact, inequality within countries rose only slightly in the nineteenth century and declined sharply in the twentieth century, as indicated by figure B6.2.2, which breaks down the Theil index between 1820 and 1992 into its international and domestic components. The figure shows that the stabilization of inequality in the middle of the twentieth century signals a turning point.

Figure B6.2.2 Decomposition of the Theil index of world income distribution. Source: Bourguignon and Morrison (2002), table 2.

Reading: For a population of N individuals of incomes $(x_1, x_2, \ldots, x_N)$, the *Theil index** is: $\frac{1}{N \ln N} \sum_{i=1}^{N} \frac{x_i}{\bar{x}} \ln\left(\frac{x_i}{\bar{x}}\right)$ where $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ is average income. Like the Gini index, it is equal to 0 for a uniform distribution and to 1 for a distribution concentrated on a single individual. It can be broken down into two components: inequality between countries and inequality within countries.

Moreover, the incidence of poverty (as defined by the World Bank, namely the proportion of the population with a real income below two dollars of 1985) continuously receded over the period. The number of poor individuals increased from 998 million in 1820 to 2.8 billion in 1992, but as a proportion of world population it decreased from 94% to 51%.

The recent accession to the middle class of a large fraction of the Chinese and Indian populations is a development of historic dimensions. However, within these two countries inequality has increased, notably between rural and urban households and across provinces, as it has for the majority of Asian countries (figure B6.2.3).

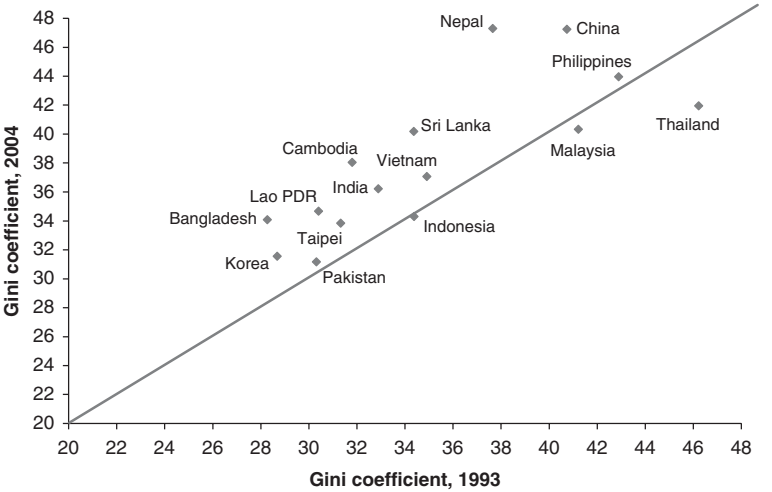


Figure B6.2.3 Inequality within Asian countries, 1994–2004.
Source: Asian Development Bank (2007, p. 32). Most data are for 1993–2004, except for Bangladesh (1991–2005), Indonesia (1993–2002), Lao PDR and Thailand (1992–2002), Nepal (1995–2003), Pakistan (1992–2004), Sri Lanka (1995–2002), and Taipei (1993–2003).

Reading: The higher the GINI coefficient, the more unequal the distribution of income within the country. Countries above the 45° line have experienced widening inequality over the period.

[a]The Gini index is equal to twice the area located between the Lorenz curve and the 45° line (cf. chapter 1). It is equal to 0 when income distribution is uniform and to 1 if all income is concentrated on only one individual.

As indicated already, causality between inequality and growth runs both ways. How income distribution affects growth is addressed in section 6.2 below. As we will discuss, this relationship is also ambiguous, since inequality can be both detrimental to growth (through restricting poor people's access to knowledge and capital) and favorable to growth (through fostering wealth accumulation).

Finally, an important issue from an equity standpoint is how global inequality has evolved. Recent research pioneered by World Bank economists allows consideration of the world distribution of income irrespective of country borders. In the nineteenth and twentieth centuries, unequal participation in the industrial revolution triggered a dramatic increase in world inequality (box 6.2). Since the 1980s, global inequality has remained stable,

as a substantial fraction of the Chinese and Indian populations have attained middle-class status while about a sixth of world population—what Paul Collier (2007) has called the "bottom billion"—remains entrenched in deep poverty.

The relation between growth and inequality is stable neither over time nor over space. This is our *fourth stylized fact.*

We finally turn to inequality within rich countries, which has become a major theme of the political debate. Technical change often encourages the hiring of skilled labor and forces job cuts in the declining sectors: It is deemed *nonneutral** or *biased** toward skilled labor and it increases income inequality.

Biased technical change is not universal: For example, the strong growth and high productivity gains of the post-World-War-II period benefited unskilled labor. However, biased technical change is not without precedent: The technological innovations brought by the Industrial Revolution provoked desperate reactions, such as the revolt of the Luddites against the new wool and cotton mills in 1811–12 in England, or the uprising of the Lyons *canuts* (silk workers) in 1831 in France.

Technical progress and growth can increase inequality within rich countries. This is the *fifth stylized fact.*

6.1.2 Catching up

Our third stylized fact is puzzling: Some countries have caught up with the most advanced ones but others have not. This deserves closer examination.

There were several episodes of convergence toward the US's level of GDP per person during the second half of the twentieth century. Western Europe first started to catch up, followed by Japan, and finally by the new industrialized countries of Asia (Hong Kong, Singapore, South Korea, and Taiwan). In Europe, Central and Eastern Europe is converging toward Western Europe. China and India are projected to reach in a few decades GDP-per-person levels comparable to those in developed countries. Is convergence a general rule? Is growth a horse race between competitors put in different starting blocks by the vagaries of history? In the long term, will all countries reach the same income level, or are they bound for different destinations?

To understand the notion of convergence, it can be useful to remember the English statistician and geneticist Sir Francis Galton, who noted in 1886 that tall men tended on average to have shorter sons and vice versa. But this "regression to mediocrity" (i.e., reversion to the average) does not mean that all men will ever have the same size within a single generation, and not even that the dispersion of sizes will diminish. The same applies for per-capita GDP. The tendency of the less-developed countries to grow faster is called *β-convergence**, because it can be measured through a positive β coefficient

in the following estimation:

$$\frac{1}{T-1}\ln\frac{Y_{iT}}{Y_{i1}} = \alpha - \beta \ln Y_{i1} + \varepsilon_{iT} \tag{6.1}$$

where Y_{i1} is the initial level of per-capita GDP of country i and Y_{iT} the final level at date T, α and β are the coefficients to be estimated and ε_{iT} is an error term. A positive β means that the lower the initial GDP per capita, the higher the growth rate.

As for men's sizes, β-convergence does not necessary imply that all countries will end up with the same GDP per capita.[18] Think, for example, of exogenous shocks due to climate, wars, the discovery of natural resources, etc. Such factors can blur the underlying convergence of per-capita GDP toward the level of the most advanced country.

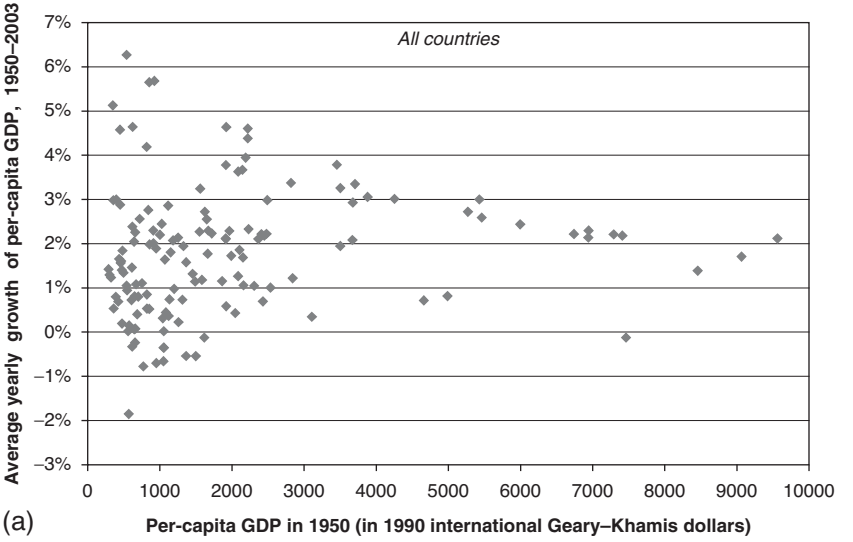
Figure 6.5 shows the link between the average growth rate of countries' per-capita GDP between the 1950-to-2003 period and its 1950 level. There is no apparent worldwide β-convergence (in plain English, poorer countries do not grow faster), but there is clearly convergence within the OECD. Such *conditional convergence** (as opposed to *absolute convergence**, or *unconditional convergence**) can be recovered by conditioning equation (6.1) on structural variables Z_{iT}:

$$\frac{1}{T-1}\ln\frac{Y_{iT}}{Y_{i1}} = \alpha - \beta \ln Y_{i1} + \gamma Z_{iT} + \varepsilon_{iT} \tag{6.2}$$

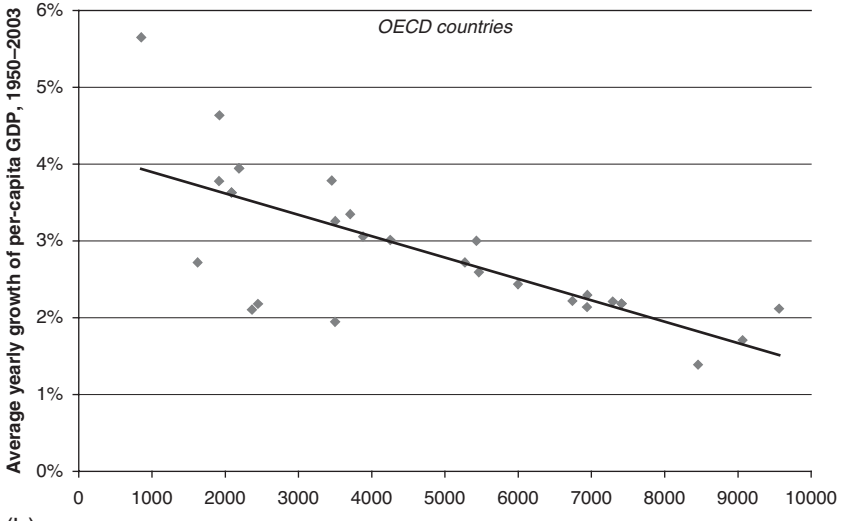
Conditional convergence means that two countries sharing the same conditioning factors converge in the long run. It implies, for example, that Serbia's GDP per person may not be able to converge toward that of Germany or France but that Slovenia is more likely to reach that goal, because Slovenia benefits from the EU legal and institutional framework while Serbia currently does not. Groups of comparable countries within which β-convergence is at play (here, the group of OECD countries) are called *convergence clubs**. A crucial issue for development-aid policies consists in understanding how any country can "join the club."

As shown by figure 6.5b, a simple regression (without conditioning variables) run on the sample of OECD countries yields a value of 1.16% a year for β, implying that half of the gap in per-capita GDP between two OECD countries would be closed within 59 years.[19] Studies that take conditional variables into account find a convergence speed of about 2.5% a year, i.e., a resorption of half the divergences in a time span of about 30 years. Conditioning variables that are usually found to have a significant long-term

18. A reduction in the dispersion of GDP per capita across countries is called σ*-convergence**.
19. Since $(1 - 0.0054)^{128} \sim 0.5$.



(a)



(b)

Figure 6.5 Catching up in the world and within OECD (β-convergence). a) All countries, b) OECD countries.
Source: Data from Maddison (2007), http://www.ggdc.net/maddison/.

Note: GDP per person in 1990, purchasing-power parity, international (Geary–Khamis) dollars.

impact on per-capita GDP are (see Barro and Sala-ì-Martin, 1995, ch. 12, for a survey):

- The quality of human capital (level of education, life expectancy);
- The functioning of markets (degree of competition, distortions introduced by state interventions, corruption);
- Macroeconomic stability (and, in particular, price stability);
- Political stability (absence of wars, coups, or frequent power shifts between opposite camps).

One problem with this approach, however, is that it implicitly assumes that the capacity to reform institutions is independent of the level of wealth. Recent research on growth has carefully tackled the reverse-causality problem, i.e., the fact that better institutions may be an outcome of growth (see below).

Empirical studies show that convergence is often *unconditional* between regions of the same country—be they US states (Barro and Sala- ì -Martin, 1991), Canadian provinces, or Japanese prefectures. In such a case, convergence tends to be unconditional because many "*Z*" factors are identical, and convergence is, moreover, encouraged by cultural homogeneity, factor mobility, and fiscal redistribution mechanisms specific to each country. Figure 6.6 provides an illustration of the US case: The downward-sloping curve clearly indicates that convergence is at work. Box 6.3 details the Canadian case.

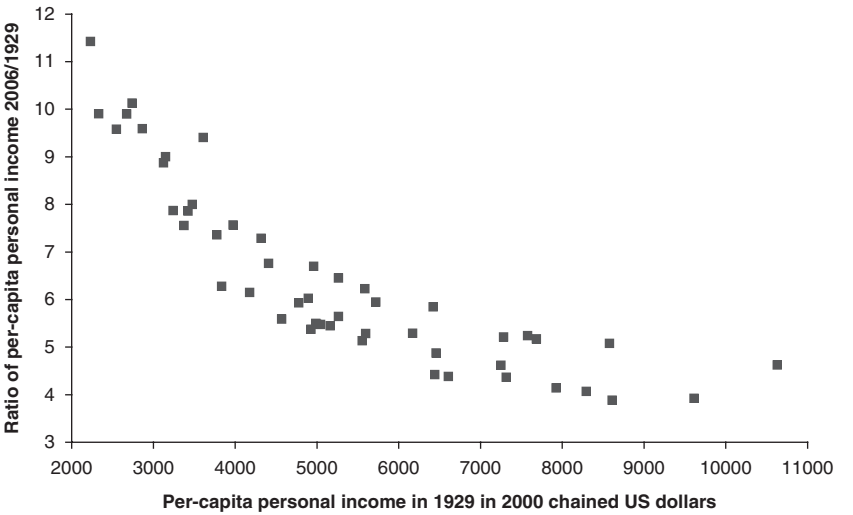


Figure 6.6 Convergence across US States.
Source: US Bureau of Economic Analysis (BEA).

Note: States' BEA data on per-person personal incomes in current US dollars have been changed into chained 2000 US dollars using the BEA data for the US GDP in current and chained 2000 US dollars.

Box 6.3 Convergence across Canadian Provinces

Figure B6.3.1 plots the per-capita incomes of the Canadian provinces net of central government transfers (a rather good proxy of GDP for which historical data is available) in 1949 and their growth rate between 1949 and 2005. The data are from Serge Coulombe, an economist with the University of Ottawa who has done extensive research on convergence (see Coulombe and Lee, 1995; Coulombe, 2007). Oil-rich Alberta is an outlier but otherwise the negative correlation between initial income and growth is almost perfect.

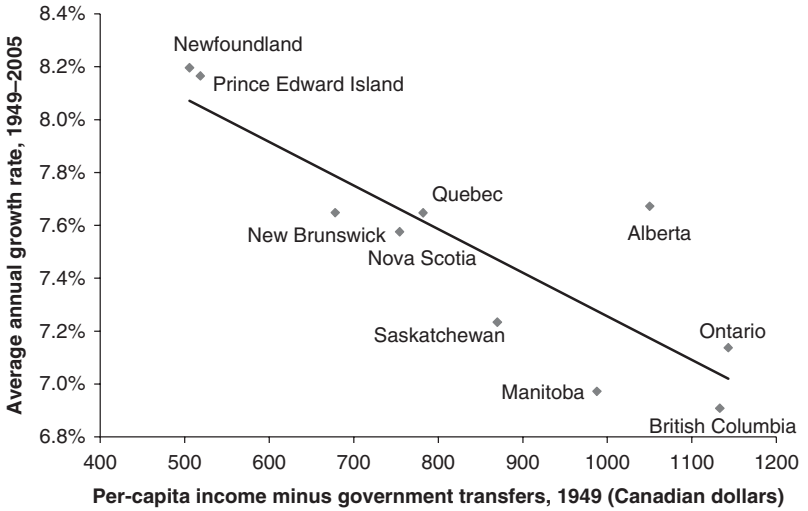


Figure B6.3.1 Convergence across Canadian provinces, 1949–2005.
Source: Data from Coulombe (2007).

Most of the convergence, in fact, took place from the 1950s to the 1980s. The remaining disparities since the mid-1980s are persistent, and can be ascribed to differences in the degree of urbanization and the resulting intensity in human capital (skilled labor moves to cities and income is consequently higher). Coulombe (2007) also finds that participation in international trade plays a role in the speed of convergence after 1980.

Persistent gaps in income per head can be found within countries. Underdevelopment in southern Italy is a case in point as is the very slow pace of convergence of the East-German regions (see below, figure 6.14 on regional GDP across Europe). Again, those can be, in part, attributed to outward migration of skilled labor. *A fortiori*, convergence between different countries is generally conditional.

6.1.3 The origin of productivity differentials

Maddison (1997) distinguishes four major determinants of long-term per-capita GDP growth: (a) Technical progress; (b) the accumulation of productive capital (i.e., of infrastructures and machines that are used for producing goods and services), which in various respects incorporates technical progress; (c) the improvement of know-how, of the level of education and of the general organization of labor; and (d) the increasing integration of the nations through trade, investment, and economic and intellectual exchange. Growth theory aims at quantifying these four determinants and at understanding their interactions and characteristics, based on rational welfare-maximizing individual behaviors.

a) A simple framework

The production function introduced in chapter 1 describes how output is produced out of capital K, labor N, and technology T:

$$Y_t = F(K_t, N_t, T_t) \tag{6.3}$$

Note that:

- Y, the goods and services output, is a *flow variable* (it is used immediately and is not transmitted from one period to another).
- K, the *capital stock*, represents the equipment and buildings available for production. It is a *stock variable* because capital is transmitted from one period to another: At time t, the economy inherits machines bought and buildings built during previous periods, and its production capacity is primarily determined by them. Each generation thus stands on the shoulders of the previous ones; this goes a long way toward explaining differences in income between countries at any point in time: Except for natural-resources producers, rich countries are primarily those with a large capital stock. The evolution of the capital stock is generally described by an equation such as $\dot{K}_t = -\delta K_t + I_t$ where δ is the rate of capital depreciation and I represents capital expenditures (a.k.a. gross fixed capital formation). Part of the capital stock thus disappears at each period (through being discarded or obsolescence) and part is renewed by the acquisition of new capital. Empirically, K is generally computed through the so-called "permanent inventory" method, i.e., by cumulating past investment flows, deflated by the replacement cost of capital, and by discarding obsolete equipment and buildings after a given lifetime.
- N, the labor input, is generally best measured by the number of hours worked, which is the product of the working-age population, the activity rate, the employment rate, and the working time (cf. box 6.1). It is also a stock (a fraction of the employed labor force leaves the market at each period, and is replaced by new entrants).

- T, the stock of technologies, is conceptually a stock as it depends on past inventions that serve in the production process before they become obsolete. However, it is more difficult to measure and less widely used.

The distinction between flows and stocks is technically useful. It is also important for economic policy, which primarily affects flows. In a context of full employment, a policy based on incentives to investment or to training thus needs to be pursued over several periods to have a significant impact on the corresponding stocks, and therefore on production.

b) Growth accounting

*Growth accounting** aims at providing a quantitative account of the role of the various determinants of growth. Its starting point is the production function, which connects real GDP Y_t at date t to factors of production: Capital K_t and labor L_t as well as to a time-varying factor linking output to the quantities of inputs used, which is called *total factor productivity** (TFP). TFP does not depend on any particular production factor (hence its name). It depends on technology, but also on the functioning of the markets and on the organization of labor.

In such a framework, the growth rate of GDP can be expressed (box 6.4) as a weighted sum of the growth rates of capital and of labor plus the growth rate of TFP, which is also called the *Solow residual** after Robert Solow (1987 Nobel Prize in economics), who introduced this decomposition (Solow, 1956). It cannot be observed and is calculated by subtraction, hence the term "residual."

Box 6.4 Growth Accounting

When specifying the production function (6.3), it is often supposed that technical progress depends on time only and affects the productivity of capital and labor symmetrically, which rewriting it as:

$$Y_t = A_t F\left(K_t, N_t\right) \tag{B6.4.1}$$

where A_t represents the effect of technical progress on the productivity of capital and labor and is called total factor productivity or TFP.[a] With constant returns to scale, an increase of TFP by x% is exactly equivalent to an increase in the quantity of capital and labor by x%.

The growth rate of income can be decomposed into the growth rates of each factor:

$$\frac{\dot{Y}}{Y} = \frac{\dot{A}}{A} + \frac{AK}{Y}\frac{\partial F}{\partial K}\frac{\dot{K}}{K} + \frac{AL}{Y}\frac{\partial F}{\partial L}\frac{\dot{N}}{N} \tag{B6.4.2}$$

where $\dot{X}$ represents the variation of the X variable, either between two consecutive periods $(X_t - X_{t-1})$ under discrete time, or as a time-derivative dX/dt under continuous time.

Defining

$$\varpi_K = \frac{AK}{Y}\frac{\partial F}{\partial K}, \varpi_N = \frac{AN}{Y}\frac{\partial F}{\partial N} \text{ and } g = \frac{\dot{A}}{A}$$

the decomposition becomes:

$$\frac{\dot{Y}}{Y} = \varpi_K \frac{\dot{K}}{K} + \varpi_N \frac{\dot{N}}{N} + g \qquad \text{(B6.4.3)}$$

The growth rate of TFP is not directly observable, but it is deduced from the above equation once Y, K, and N are known. It is the so-called Solow residual. Several methods can be used to calculate the Solow residual.

Method 1. Let us denote by c^K the *user cost of capital**, which represents the real cost of using of a unit of capital during the period of production;[b] and w the real wage (i.e., the nominal wage divided by the price level). In a competitive economy, factors' costs are equal to their *marginal product**,[c] so that $c^K = A\partial F/\partial K$ and $w = A\partial F/\partial N$. ϖ_K and ϖ_N are, therefore, the respective shares of capital and labor earnings, $c^K K$ and wN respectively in the firms' income. They can be observed from national accounts (roughly, at a country level, $\varpi_N \sim 0.6$ and $\varpi_K \sim 0.4$). In a closed, competitive economy and under constant returns to scale, $\varpi_K + \varpi_N = 1$. The "Solow residual" g is then deduced from the previous equation.

Method 2. An econometric rather than accounting method consists in regressing $\frac{\dot{Y}}{Y}$ on $\frac{\dot{K}}{K}$ and $\frac{\dot{N}}{N}$ and in extracting the residual g. Coefficients ϖ_N and ϖ_K are thus estimated, rather than calibrated. This method is delicate to implement because K and N are measured with an error and because they are correlated in the short run with the dependent variable Y, which requires the use of instrumental variables (see box 6.14).

Method 3. From the above equations, under constant returns to scale and under the hypothesis that factor incomes are equal to their marginal product, it can be shown that:

$$g = \varpi_K \frac{\dot{c}^K}{c^K} + \varpi_N \frac{\dot{w}}{w} \qquad \text{(B6.4.4)}$$

The marginal increase of TFP is therefore given back to workers and capital-owners.

These methods can be generalized when more than two factors are used in production; for example, when energy consumption is taken into account or when capital and labor are broken down into several categories. This breakdown is useful to limit biases in the calculation of g. Indeed, the categories of capital and of labor that develop fastest are also those whose return increases fastest: High stock market yield of new technologies, wage rises in peak sectors. If the corresponding rise of the volume of K

or of N were weighted by the average return or the average wage in the economy, their contribution to g would be underestimated.

The first and the third approaches are valid only if the factor returns are equal to their marginal products. This is neither the case in the presence of distortions, nor in the presence of externalities, because the private marginal product then differs from the social marginal product.

[a] This is a strong assumption known as the *Hicks-neutrality** of technical progress. Alternative assumptions are the so-called *Solow neutrality**, which assumes that technical progress is equivalent to an augmentation of the quantity of labor used in production (and therefore increases the marginal productivity of capital) and the so-called *Harrod-neutrality**, which is based on the opposite assumption that technical progress is equivalent to an augmentation of the quantity of capital used in production. With Solow-neutral technical progress the production function can be written as $Y_t = F(K_t, A_t N_t)$, while with Harrod-neutral technical progress the production function is can be written as $Y_t = F(A_t K_t, N_t)$.

[b] Capital is acquired at a certain price; it can be resold at the end of the period with a discount corresponding to its depreciation; moreover, financing the investment carries a rate of interest (or, equivalently, an opportunity cost), and the corresponding interests must be paid. On the whole, the user cost of capital is equal to the real interest rate plus the rate of capital depreciation.

[c] The marginal product of a factor is equal to the output that an additional unit of this factor produces, all other factor quantities remaining constant.

c) Labor productivity versus total factor productivity

TFP should not be confused with labor productivity Y/N. Noting per-capita capital $k = K/N$, and under constant returns to scale (box 6.4), the growth rate of labor productivity over time can be decomposed into two components:

$$\frac{\dot{Y}}{Y} - \frac{\dot{N}}{N} = g + \varpi_K \dot{k} \tag{6.4}$$

The first term g, or TFP growth, represents technical progress in the broad sense (as it also encompasses institutional factors). The second term is the growth rate of the capital stock per person, i.e., the growth rate of the *capital intensity** of the production process. This decomposition helps understanding that labor productivity growth can come either from an acceleration of TFP or from an increase in capital intensity (also called *capital deepening**), i.e., from a substitution of capital to labor in the production process.

Let us, for example, look at the origin of the growth gap between Europe and the US (table 6.1). The annual growth rate of labor productivity (per hour) in the US increased from 1.2% over the period 1990–95 to 2.3% during the so-called "new economy" period of 1995–2000.[20] Given the 0.6% per

20. The *new economy** was a paradigm developed in the US in the late 1990s to designate a high-growth, low-inflation regime supported by the productivity-improving spread of information and communication technologies.

Table 6.1
Growth accounting in the US and in the EU (average annual growth rates, in %)

	US			EU (15)			Gap (US – EU)		
	1990–95	1995–2000	2000–04	1990–95	1995–2000	2000–04	1990–95	1995–2000	2000–04
GDP (**1**)	2.5	4.2	2.4	1.6	2.7	1.5	0.9	1.5	0.9
Total hours worked: (**2**) = (**3**) + (**4**)	1.3	1.9	−0.4	−0.9	0.9	0.4	2.2	1.0	−0.8
Employment (**3**)	1.1	1.7	0.4	−0.5	1.4	0.7	1.6	0.3	−0.3
Working hours (**4**)	0.2	0.2	−0.8	−0.4	−0.5	−0.3	0.6	0.7	−0.5
Labor productivity: (**5**) = (**1**) − (**2**)	1.2	2.3	2.8	2.5	1.8	1.1	−1.3	0.5	1.7
Contribution of capital/labor ratio (**6**)	0.7	1.2	1.1	1.3	0.9	0.7	−0.6	0.3	0.4
TFP: (**7**) = (**5**) − (**6**)	0.5	1.1	1.7	1.2	0.9	0.4	−0.7	0.2	1.3

Source: Data from Timmer, Ypma and Van Ark (2003, Appendix tables, updated 2005). Groningen Growth and Development Centre.

year acceleration in the number of hours worked (from an average yearly growth rate of 1.3% to 1.9%), the growth rate of GDP increased by 1.7 percentage points per year between the two periods (from 2.5% to 4.2%). The acceleration of labor productivity was half due to TFP (as a consequence of the internal restructuring of firms and the introduction of information and communication technologies) and half to intensified capital deepening (as the result of a massive investment since the beginning of the 1990s). GDP growth thereafter slowed down in 2000–04 (to 2.4% per year): While TFP growth accelerated further to 1.7 per year on average, capital deepening remained constant and the total number of hours worked declined at a rate of –0.4% per year over the period.

European growth also accelerated, by more than a full percentage point per year (from 1.6% to 2.7%) from 1990–95 to 1995–2000, but mainly through faster employment growth (+1.9% between the two periods), and despite both a decline in capital intensity (–0.4%) and a TFP slowdown (–0.3%). TFP decelerated further between 2000 and 2004.

To sum up, the US and Europe have traded places. In the early 1990s (as in the previous decade) the US was a slow-productivity, high-employment growth economy while the growth pattern in the EU was the opposite. By 2000, the US had become a high-productivity, slow-employment growth economy while the EU had adopted a growth pattern resembling that of the US in the 1980s.

However, the 2007–09 economic crisis will have a lasting, if as yet uncertain, impact on productivity and potential output for OECD countries. There are three main channels through which the level of potential output might be permanently reduced. A first channel goes through the labor market: Hysteresis effects due to the rise in unemployment might raise structural unemployment (see section 6.3 below) and labor force participation might be reduced. This impact might be partly offset by migration flows in some countries. A second channel takes place through a lower capital–labor ratio due to a higher cost of capital, after a long period of unusually low real long-term interest rates. A third channel, namely the evolution of TFP, remains ambiguous, as policy responses might prevent R&D expenditures from being cut and might promote human capital accumulation, and as resources may be shifted from less-efficient to more-productive activities ("creative destruction," see section 6.2 below). Overall, the OECD (2010) estimates that the crisis may lead to a 3.5% medium-term cut to the pre-crisis level of potential output for the OECD area as a whole, and to a 4–4.5% cut for a typical average-sized OECD country (for a peak loss, around 2013, of about 3% for the US, 5% for Greece, 9% for Spain). These reductions in potential output would be due, in equal proportions, to the first two channels mentioned above, the last channel remaining ambiguous. However, the crisis would not affect the rate of potential growth (as opposed to the level of potential output) in the longer term. Potential growth is expected to decelerate for other reasons, notably population ageing.

Measures of labor productivity components are of course statistical constructs, and it remains difficult to quantify the respective roles of technical progress and of capital intensity. A telling debate took place in the 1990s on the sources of the Asian "miracle" (at that time the growth of small East Asian economies). According to Alwyn Young (1992), the astonishingly strong growth of those since the 1960s involved nothing miraculous, but was due to massive capital accumulation encouraged by "Colbertist" policies[21] (very low rates of interest, proactive industrial policy, etc.) and not to total factor productivity (box 6.5). In short, as Paul Krugman later wrote (1994b), the Asian miracle was "the fruit of perspiration, not of inspiration." As capital and labor exhibited decreasing returns, growth ineluctably had to slow down. The 1997–98 financial crises partially validated the diagnosis of Young and Krugman, since they were, inter alia, a consequence of over-investment.

Box 6.5 A Tale of Two Cities: Asian Growth According to Alwyn Young

In 1992, MIT economist Alwyn Young compared the economic models of Hong Kong and Singapore. Both cities had similar histories as UK enclaves in the Chinese world, with large commercial ports, having developed their manufacturing industry after World War II and then financial services. The levels of their per-capita GDP were identical in the 1960s, and their rates of growth were comparable between 1960 and 1990. But the resemblance stopped there. After a careful growth-accounting study, Young concluded that the growth in Singapore came primarily from the accumulation of productive capital, while total factor productivity slowed down. Singapore was a "victim of its own targeting policies, which are increasingly driving the economy ahead of its learning maturity into the production of goods in which it has lower and lower productivity." (Young, 1992, p. 16). In 1994, Paul Krugman (1994b) insisted that: "There is no sign at all of increased efficiency. In this sense, the growth of Lee Kuan Yew's Singapore is an economic twin of Stalin's Soviet Union growth achieved purely through mobilization of resources." In contrast, Hong Kong could maintain a rapid progression of total factor productivity. Young explained this contrast by very different growth models: Free-market in Hong Kong, central planning in Singapore. In another article, Young (1995) extended his conclusions to other Asian "dragons": Growth there was due to capital accumulation, to the labor force and to education, but not to technical progress.

21. Jean-Baptiste Colbert was in the seventeenth century a minister under French king Louis XIV and the architect of its economic policy. What became known as *Colbertism** involves systematic state intervention in the development of supply and the promotion of exports.

Both the figures and the diagnosis were sharply criticized by other economists who, on the basis of different measures of the share of factors in value added, constructed a rate of TFP growth in Singapore[a] higher than Young's.

[a] See Iwata et al. (2003) for a synthesis of this debate.

d) Where do productivity gains come from?

Growth accounting reaches its limits with the opacity of the Solow residual. Solow himself (1956) found that g explained some seven-eighths of the doubling of labor productivity in the US from 1909 to 1949, while the increase in capital intensity only explained the remaining eighth. To understand TFP divergences, a closer analysis is needed.

A first approach consists in relating TFP growth to indicators of human capital development such as: The literacy rate or the number of graduates by age group; innovation such as the share of research and development (R&D) expenditures in GDP; or technological development such as the penetration of computers, the number and speed of Internet connections. This, for example, is what underlies frequent cross-country comparisons of R&D expenditures, the usual praise of Japan's effort, and the EU's efforts to increase its R&D ratio from about 2% to 3% of GDP.

Macroeconomic approaches through growth accounting and measures of the innovation effort at a microeconomic level are not easily connected, however. For a long time, computers have been seen everywhere—except in the productivity figures (to paraphrase Robert Solow's famous 1987 statement).[22] This *Solow paradox** probably disappeared in the second half of the 1990s, with the advent of the new economy. TFP clearly accelerated, at least in the US, through the development of Information and Communication Technologies (ICT). Figure 6.7 compares investment in the ICT-producing industry branches (computer hardware, software, communications hardware) and highlights that Europe (particularly continental European) lags behind the US.

The contribution of ICT to growth takes place through several channels.[23] Through substituting capital for labor, managing inventories and making better use of inputs, ICT raises TFP. It also induces reallocations within the labor force and frequently the substitution of a capital-skilled labor mix for unskilled labor.

However, the impact of new technologies can be delayed, particularly because the full effect on productivity requires complementary investments in

22. "You can see the computer age everywhere but in the productivity statistics," *New York Review of Books*, 12 July 1987.
23. See OECD (2003) for a summary.

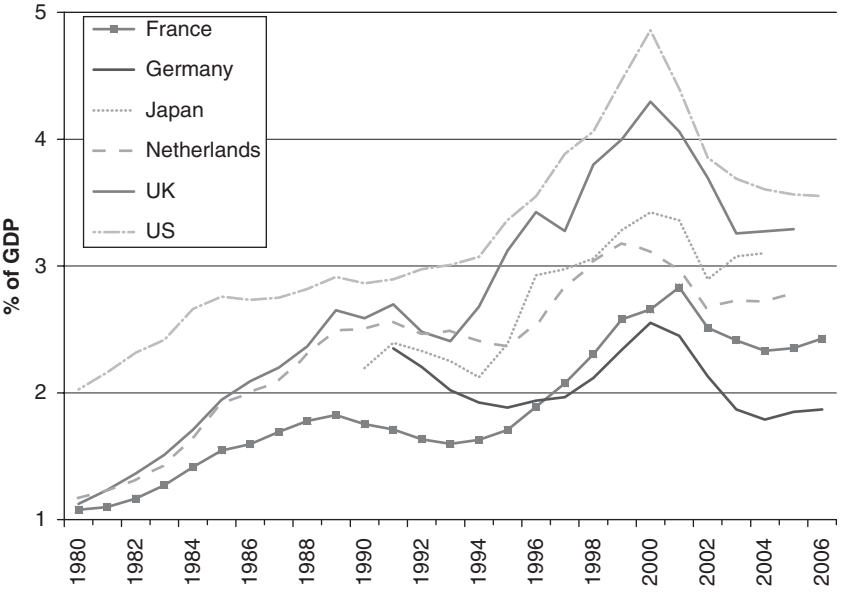


Figure 6.7 ICT investment rates in % of GDP.
Source: OECD.

other forms of capital (for example, firms need to reorganize in order to use the new ICT equipment as much and as efficiently as possible). This provides one of the explanations of the Solow paradox, in line with the interpretations of the historian of innovation Paul David. David (1990) notably highlighted that it had taken a long time for the invention of electricity to affect productivity. Such effects have been documented in particular by Basu et al. (2004), Yang and Brynjolfsson (2001), or, in France, by Askenazy (2001).

The ICT volume is poorly apprehended by national accounts because equipment and software prices have been decreasing rapidly. National accountants have developed *hedonic price indices** that take account of the quality improvements brought by the new generations of products. Instead of focusing on the price of the product itself, hedonic prices are based on the services delivered through the product (for example, in the case of computers, the hedonic price would take into account the memory capacity, processing speed, screen resolution, etc.). On the whole, the contribution of ICT to GDP growth during the second half of the 1990s is estimated at about one percentage point a year in the EU and 1.7 percentage points in the US, 0.6 and 1 points respectively being related to the substitution of capital for labor, and 0.4 and 0.7 to TFP growth (table 6.2). From 2000 to 2004, ICT contributed 0.5 percentage points of GDP growth in the EU, and 0.9 points in the US.

What are the other sources of productivity gains beyond ICT? The diffusion of technologies is difficult to track. Robert Gordon (2000), a renowned specialist in productivity and growth, first claimed that the acceleration of

Table 6.2
ICT contribution to GDP growth

	US		EU (15)		Gap (US – EU)	
	1995–2000	2000–04	1995–2000	2000–04	1995–2000	2000–2004
Labor productivity (**1**)	2.3	2.8	1.8	1.1	0.5	1.7
Contribution of capital deepening (**2**)	1.2	1.1	1.0	0.8	0.2	0.3
ICT	1.0	0.6	0.6	0.3	0.4	0.3
Non-ICT	0.2	0.5	0.4	0.5	−0.2	0.0
TFP (**3**) = (**2**) − (**1**)	1.1	1.7	0.9	0.4	0.2	1.3
ICT	0.7	0.3	0.4	0.2	0.5	0.1
Non-ICT	0.4	1.4	0.5	0.2	0.0	1.2

Note: ICT and capital deepening data are taken from table 2 in Van Ark and Inklaar (2005).

Source: Groningen Growth and Development Centre.

TFP in the US was circumscribed in the computer sector and, apart from that sector, was primarily cyclical (because it corresponded to a period of expanding demand), before revising his assessment.[24] Van Ark and Bartelsman (2004) identified the five sectors that have most contributed to US productivity over the period 1995–2002: They include retail trade, wholesale trade, electronic components, financial intermediation and its ancillary services. These five sectors explain more than half the gains in US productivity over the period. In comparison, the five best-performing sectors in Europe (communications, information technology, legal services and publicity, electronic components, and social and health services) explain less than half the gains in European productivity; moreover, the five best-performing sectors in the US grew twice as fast. In summary, the most dynamic US sectors have grown much faster than the most dynamic European sectors. The 2007–09 crisis, however, questions the sustainability of past US dynamism.

6.2 Theories

Growth accounting is a description, not an explanation of economic growth. In order to understand its mechanisms and assess the role of economic policy, we need to turn to theory and to investigate the determinants of labor-force growth, of capital accumulation, and of technological innovation.

In pre-industrial theories like those of Thomas Malthus (1798), fertility was indeed regarded as the fundamental determinant of growth. According to Malthus, it adjusted to technological shocks so that the standard of living remained constant. Subsequent theories generally included demography as an exogenous phenomenon, except sometimes for migrations (cf. Barro and Sala-ì-Martin, 1995, ch. 9). They focused mainly on capital accumulation. In the 1980s, they had reached, however, a somewhat frustrating conclusion: Under decreasing marginal returns to capital, incentives to accumulate capital fade over time, so that any growth in GDP per person can only stem from a constant flow of technological innovation. In the 1960s and 1970s, very few economists remained interested in growth theory, and in their review of the literature, Hahn and Matthews (1964, p. 890) even concluded that it might have reached the point of diminishing returns.

A renewal started in the late 1980s with the advent of the so-called endogenous growth theory, which focuses on the determinants of total factor productivity. In this context, economists revisited Joseph Schumpeter's seminal ideas on what drives innovation, and ventured into new areas such as the interaction between growth and geography or between growth and the quality of institutions. Standard models were also revisited to shed light on economic development. Growth theory has thus nowadays become one of the most active branches of economic analysis.

24. See Gordon (2000, 2003), Oliner and Sichel (2002), and, in the French case, Cette et al. (2004).

In what follows, we first present the standard models of growth through capital accumulation with exogenous technical progress; we then turn to models with endogenous technical progress; finally, we discuss the role of deep growth determinants such as geography, income distribution, and institutions.

6.2.1 Growth through capital accumulation

The basic tool for the analysis of growth is the production function of section 6.1, which we assume can be written:

$$Y_t = A_t F(K_t,\ N_t) \tag{6.5}$$

where Y denotes output, A is technical progress, K is the capital stock, and N is the labor force (or the total number of hours worked). As growth analysis deals with medium-to-long-term horizons, it is generally assumed that the economy is at full employment. Therefore, employment N is equal to the active labor L, so that:

$$Y_t = A_t F(K_t,\ L_t) \tag{6.6}$$

The assumption of full employment may seem to run counter to experience. However, this specification also accounts for situations of persistent unemployment. Let us suppose that there is a rate of structural unemployment u that cannot be permanently reduced whatever the level of aggregate demand. The above formulation can be adapted by replacing L by $L(1 - u)$.[25] Hence we can use it without implying any assumption about the structure of the labor market.

a) First steps: growth and disequilibrium

Theories of the accumulation of productive capital in a closed economy have as a starting point the equilibrium between the supply of capital, i.e., the flow of savings, and the demand for capital, i.e., the flow of investment desired by profit-maximizing companies.

A first intuition was developed independently by economists Roy Harrod in 1939 and Evsey Domar in 1946, who highlighted the risk of economic instability in the growth process. They pointed out that the growth rate of the capital stock determined by investment (and therefore by the savings rate) does not spontaneously correspond to the growth rate that is necessary to maintain full employment. They, therefore, saw a risk either of a shortage of labor (leading to inflation) or of a shortage of capital (leading to unemployment): Balanced growth was possible only if, by sheer coincidence, the economy

25. However, this formulation ignores any interdependence between short-term fluctuations and long-term growth, such as hysteresis in the rate of unemployment. This is a point to which we return at the beginning of section 6.3.

remained on the "*razor's edge*"* in which the savings–investment balance corresponds to the full employment equilibrium (box 6.6).

The *Harrod–Domar model** does not provide a realistic description of long-term growth. Domar himself considered it rather as a study of the interaction between temporary shortages of demand and investment, in the context of the consequences of the 1929 crisis and then of the war economy characterized by a shortage of capital. A model which predicts that growth is constrained by a shortage of capital also provided a description of Europe in the immediate post-World-War-II period, and a theoretical justification for the reconstruction-aid policies meant to compensate for the European countries' insufficient savings and fill their "financing gap." This approach was explicit in the June 1947 *Marshall plan**,[26] and even more in the "national assessment" simultaneously prepared in France by Planning Commissioner Jean Monnet.

Box 6.6 The Harrod–Domar Model

The model developed by Harrod (1939) and Domar (1946) assumes a production function with *complementary inputs**:

$$Y_t = min\ (AK_t,\ BL_t) \qquad \text{(B6.6.1)}$$

where Y_t is output, K_t is the stock of capital, A and B are constant parameters and the labor force L_t grows at a constant rate n. This kind of formalization corresponds to the assumption that technology is fixed and that efficient production requires capital and labor inputs to be in a constant proportion.

As inputs are complementary, full employment requires a sufficient capital stock to employ all the labor force. But if the capital stock is higher, production capacity will not be fully employed due to a shortage of labor. However, labor growth and the growth of the capital stock are determined independently. Full employment equilibrium therefore hinges on a stroke of luck.

The evolution of the capital stock over time is:

$$\dot{K}_t = -\delta K_t + I_t \qquad \text{(B6.6.2)}$$

where δ is the rate of capital depreciation and I_t is gross fixed capital formation (investment). Investment is financed by available

26. The European Recovery Program, better known as the "Marshall plan" (after the American Secretary of State George C. Marshall) was a program of financial assistance by the US for rebuilding the countries of Europe; it cost overall $13 billion over four years, which represented 5.3% of the 1947 US GDP. The USSR was invited to take part but refused. The institution set up to implement the plan, called the Organization for European Economic Co-operation (OEEC), later became in 1960 the *Organization for Economic Cooperation and Development** (OECD).

savings: $I_t = \sigma Y_t$ where s is the constant savings rate. As long as the capital stock remains low, its evolution is given by:

$$\dot{K}_t = (\sigma A - \delta)K_t \qquad \text{(B6.6.3)}$$

The trajectory of the economy depends on the level of the savings rate: If $\sigma > \delta/A$, then the capital stock and output grow at the constant rate $\sigma A - \delta$, until $AK_t = BL_t$, after which output growth is limited by the availability of labor. Below a certain threshold for the capital/labor ratio, a policy favorable to saving increases the growth rate.

As long as the capital stock remains low, the production function is $Y_t = AK_t$: This is why growth models with constant returns to capital are known under the generic name of "AK models."

b) Saving, investment and balanced growth

Firms, however, do not invest in order to use available savings but in order to make profits: The return to capital is the main engine of investment. In 1956, recognition of this microeconomic incentive led Robert Solow and Trevor Swan to separately develop a model that has carried considerable intellectual influence and still provides a reference framework for the analysis of economic growth.

Unlike Harrod and Domar, Solow and Swan describe a growth path where markets are in balance. Production factors are *substitutable**—therefore, there are no more "razor's edge" equilibriums—and the marginal return to capital is decreasing. The more capital is accumulated, the less profitable it is at the margin, and the incentive to invest vanishes when the marginal return on capital is equal to the user cost of capital, i.e., when adding to the capital stock costs exactly the value of the additional production it brings. At this stage, the per-capita level of the capital stock, and therefore also (under constant returns to scale) per-capita GDP, are stable over time (see the detailed model in box 6.7). The corresponding growth path is called the *steady state**.

In the basic *Solow–Swan model** the savings rate is exogenous, production exhibits constant returns to scale, labor and capital are perfect substitutes, and each exhibits decreasing marginal returns. There is an equilibrium value of the per-capita level of the capital stock k^* that only depends on the savings rate and on the rate of capital depreciation. In a more complete version where population and TFP grow over time at respective exogenous and constant rates n and g, it is the TFP-adjusted level of per-capita GDP that is stable in the long run, and its value also depends on g. The capital stock and GDP both grow at a constant rate $n + g$, and, under this specification, the model is fully consistent with growth accounting as introduced above.

In this category of models, when the capital stock reaches its equilibrium value, the *growth rate* of GDP only depends on demography and on exogenous

technical progress (*not* on the savings rate, which determines the level of GDP per person but not its rate of growth along the stationary path). Growth is temporarily faster when the economy starts from an initial situation of capital shortage (which explains accelerated catching up by developing economies), but it sooner or later adjusts to the (lower) steady state. Hence a first, disappointing, conclusion for economic policy: Under decreasing marginal returns to capital, policies aimed at encouraging saving or investment are not able to influence the long-term *growth rate*, but only the long-run *GDP-per-capita level*.

Box 6.7 The Solow–Swan Model

The production function is of the Cobb–Douglas type: $Y_t = AK_t^{\alpha} L_t^{1-\alpha}$ with $0 < \alpha < 1$, which implies both decreasing returns to each production factor taken separately and constant returns to scale. Labor and capital are perfect substitutes. Assume first that total factor productivity A is constant. In a closed economy, output Y is equal to the income distributed to economic agents. Labor supply grows at a constant rate n. The capital stock increases every year by the volume of investment I, but in each year a fraction δ of it is discarded. In this closed economy, a fraction σ of income is saved and invested every year. Thus:

$$\dot{L}_t = nL_t \qquad \dot{K}_t = -\delta K_t + I_t \qquad I_t = \sigma Y_t \tag{B6.7.1}$$

Let lower-case letters represent variables per person:

$$k_t = K_t/L_t \qquad y_t = Y_t/L_t \qquad \sigma_t = \sigma Y_t/L_t$$

The dynamics of k_t are given by:

$$\frac{\dot{k}_t}{k_t} = \frac{\dot{K}_t}{K_t} - \frac{\dot{L}_t}{L_t} = -\delta + \sigma \frac{y_1}{k_t} - n \tag{B6.7.2}$$

On the steady-state trajectory where variables per person Y^*, k^*, and the level of savings per person s^* are constant, capital accumulation is determined by the following equation:

$$\sigma y^* = (n + \delta)k^* \quad \text{with} \quad y^* = k^{*\alpha} \tag{B6.7.3}$$

This means that, in the steady state, the savings of each period exactly finance the capital expenditure necessary to replace the depreciated capital and to equip the new workers. The stock of capital per person thus remains constant. The steady state levels of per person output, capital, and savings are written as:

$$y^* = \sigma^{\frac{\alpha}{1-\alpha}} (n + \delta)^{-\frac{\alpha}{1-\alpha}} \qquad k^* = \sigma^{\frac{1}{1-\alpha}} (n + \delta)^{-\frac{1}{1-\alpha}}$$

$$s^* = \sigma^{\frac{1}{1-\alpha}} (n + \delta)^{-\frac{\alpha}{1-\alpha}} \tag{B6.7.4}$$

In figure B6.7.1, where the X-axis represents the level of the capital stock per person and the Y-axis represents savings and investments per person, the steady-state equilibrium corresponds to the intersection of the curve $\sigma_t = \sigma k_t^{\alpha}$ (representing savings), and of the straight line $(n + \delta)k_t$, which represents the investment per person necessary to maintain k_t constant:

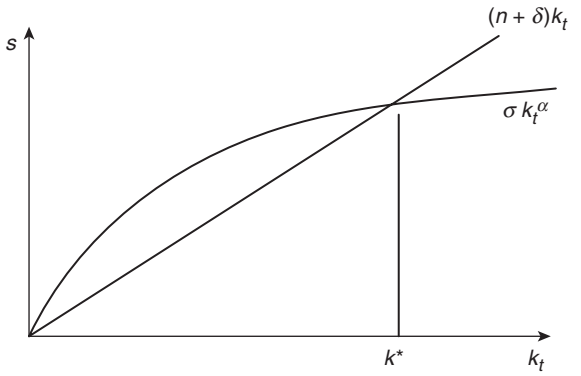


Figure B6.7.1 The Solow–Swan model.

The steady-state equilibrium is stable: Whatever the initial value k_0, the capital stock per person tends toward k^* when t tends to infinity. The model therefore leads to the following two conclusions:

- In the long run, the levels of the capital stock per person and of income per person are constant. Income grows at a constant rate that depends only on demography;
- In the long run, income per person depends positively on the savings rate, all the more that capital plays an important role in the production function (in a competitive economy, α represents the share of capital income and $(1 - \alpha)$ the share of labor income in output).

The first conclusion is disappointing: The model does not account for the fact that income per person grows over time. The only possible explanation in this model is that total factor productivity increases over time. Let us now suppose that total factor productivity A grows at rate g:

$$\dot{A}_t = gA_t \tag{B6.7.5}$$

and let us again solve the model for the steady state. Results obtained are similar, but for the inclusion of trend growth rate. We find:

$$y^* = \sigma^{\frac{\alpha}{1-\alpha}}(n + g + \delta)^{-\frac{\alpha}{1-\alpha}} \qquad k^* = \sigma^{\frac{1}{1-\alpha}}(n + g + \delta)^{-\frac{1}{1-\alpha}}$$

$$s^* = \sigma^{\frac{1}{1-\alpha}}(n + g + \delta)^{-\frac{\alpha}{1-\alpha}} \tag{B6.7.6}$$

where lower-case letters now represent variables per "effective labor unit": $y = Y/AL$ and $k = K/AL$.

The model predicts that, all other things being equal, a 1% increase of the savings rate leads to an $\alpha/(1-\alpha)$% increase of steady-state per-capita GDP. Per-capita income y and the per-capita capital k grow at rate g, and income grows at rate $n + g$ over time. The parameter g can be interpreted as measuring the pace of technical progress. But it is assumed to be exogenous: The model is silent about its origin.

Up to now, we have not introduced any normative assumption but we have simply drawn logical consequences from the assumption of a decreasing marginal return to capital. We can now use the model for a normative purpose. The social objective cannot be to reach the highest possible level of per-capita GDP. Indeed, as box 6.7 explains, this requires maintaining a high per-capita capital stock, which necessitates allocating to capital accumulation a large fraction of income that is, therefore, not available for consumption and does not contribute to the individuals' immediate well-being. From a normative standpoint, GDP per person should therefore be high enough to make resources available, but not too high, otherwise replacement investment would absorb too large a share of GDP.[27]

This suggests that there might be an "optimum" level of the per-capita capital stock and therefore of per-capita GDP, a question addressed as early as in 1928 by Frank Ramsey. He assumed that the social objective was to maximize *per-capita consumption* on a sustainable basis. Let us suppose that a benevolent planner (cf. chapter 1) can choose the households' savings rate. The *Ramsey model** shows (box 6.8) that a savings rate exists that maximizes per-capita consumption. With a Cobb–Douglas production function, this optimal savings rate is exactly equal to the weight of capital in the production function. It is as if capital income (i.e., dividends paid by firms) were entirely reinvested in the economy, while labor income was consumed. At the optimum, the model shows that the marginal return to capital (i.e., the real interest rate) is exactly equal to the GDP growth rate, $n + g$: This relation is called the *golden rule** of capital accumulation. When it is verified, a marginal increase of the capital stock generates an additional income that exactly covers the additional expenditure needed to maintain that additional unit of capital, so that per-capita consumption remains unchanged.

The golden rule provides a simple means for identifying the optimal growth trajectories: If the interest rate is durably higher than the growth rate, there is "not enough" capital, and a higher savings rate would allow raising consumption (consumption would have first to decrease to give way to an

27. Incidentally, this shows why GDP maximization cannot be taken as a criterion for evaluating policies.

increase in savings, but it would eventually benefit from the consecutive rise in per-capita GDP); if the rate of interest is durably lower than the growth rate, there is "too much" capital and citizens would be better-off using the income from it for consumption rather than investment. The former situation can be found in developing economies where too much of the income is consumed, while the latter is called *dynamic inefficiency**, and is found in economies where incentives are distorted in favor of investment, such as China or, as described by Young, Singapore (box 6.5). Such considerations may play an important role in deciding whether pension schemes should be funded and invested in the economy, as explained in section 6.3.

Box 6.8 The Ramsey Model and the Golden Rule

Let us start with the model of box 6.7 without technical progress ($g = 0$). The government is assumed to choose the savings rate (for example, by means of tax measures, cf. chapter 7) so as to maximize long-term per-capita consumption $(1 - \sigma)k^{*\alpha}$. The optimal savings rate comes out as:

$$\hat{\sigma} = Arg \max c^*(\sigma) = Arg \max \left[(1-\sigma)\sigma^{\frac{\alpha}{1-\alpha}}(n+\delta)^{-\frac{\alpha}{1-\alpha}}\right] \quad \text{(B6.8.1)}$$

Caps on per-capita variables designate variables along the optimal growth trajectory. Simple calculations show that $\hat{\sigma} = \alpha$ (beware that this simple result holds only when the production function is Cobb–Douglas). The optimal growth trajectory has an interesting property. From the results of box 6.7, the *marginal productivity of capital** on this trajectory is:

$$\frac{\partial y}{\partial k} = \alpha \hat{k}^{\alpha-1} = \frac{\alpha \hat{y}}{\hat{k}} = n + \delta \quad \text{(B6.8.2)}$$

However, profit maximization in a competitive environment implies that this marginal productivity is exactly equal to the user cost of capital c^k so that:

$$\frac{\partial y}{\partial k} = c^k = r + \delta \quad \text{(B6.8.3)}$$

where r is the real interest rate. These two relations imply that on the steady-state path that maximizes per-capita consumption, the real interest r rate is equal to the growth rate of the economy n. This is the so-called *golden rule*. This result also applies when $g \neq 0$, in which case, $r = n + g$.

The result can be also represented in figure B6.8.1, which, like in box 6.7, represents per-capita income and per-capita savings (equal to per-capita investment) as functions of the stock of capital per person. At the steady state, and whatever the value of σ, investment σk^α is exactly equal to the increase in the capital stock needed to maintain the stock of capital constant, namely $(n + \delta)k$, which leads to the value of the

steady-state per-capita capital stock k^*. Per-capita consumption is represented in the figure by the distance between the two curves k^α and σk^α. The figure shows that this distance is maximum at $k = \hat{k}$ where the tangent to the production function $y = k^\alpha$ is parallel to the line $(n + \delta)k$. This leads to the golden rule $r = n$, since the marginal productivity of capital $\frac{\partial y}{\partial k}$ is equal to the user cost of capital $r + \delta$. In the figure below $\sigma > \alpha$ and therefore $k^* > \hat{k}$: There is "too much" savings, "too much" capital, and the real interest rate (measured by the tangent in k^* to the curve $y = k^\alpha$) is lower than the growth rate (measured by the tangent in $\hat{k}$, $n+\delta$).

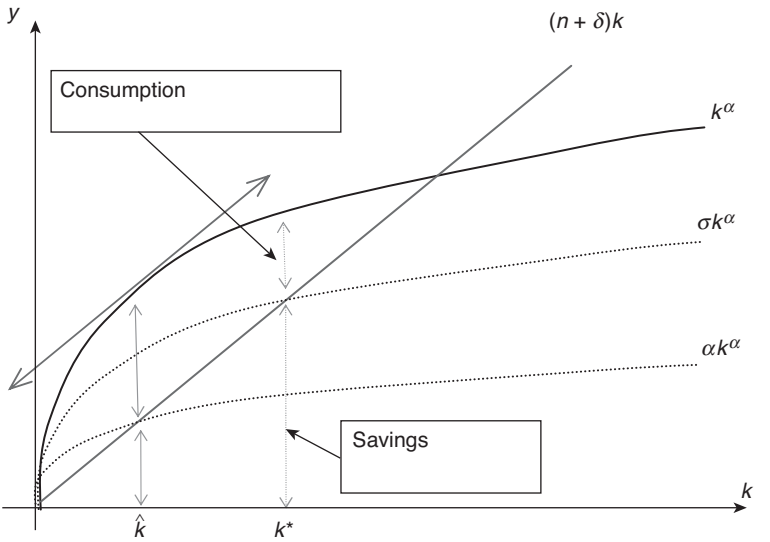


Figure B6.8.1 Consumption and savings at the social optimum.

c) Growth and catching-up

The Solow and Swan model provides a theoretical framework for growth accounting whose empirical importance was emphasized in a previous section. But are its assumptions realistic? With a Cobb–Douglas production function, the model predicts (box 6.7) that a one-percentage-point increase of the savings rate, other things being equal, leads to a $\alpha/(1-\alpha)$% increase in per-capita GDP, where α is the weight of capital in the production function. In 1992, N.G. Mankiw, D. Romer and D.N. Weil tested this relation on a panel of countries and found an elasticity of per-capita GDP to the savings rate of approximately 1.5, consistent with a value $\hat{\alpha}$ of 0.6: In a closed economy under constant returns to scale, capital income would thus absorb 60% of the value added! In reality, α is known to be close to 30–40%.

Mankiw et al. propose an explanation: TFP is not exogenous and instead depends on the accumulation of a second type of capital, namely *human capital**. Indeed, a part of national savings is invested in education and training and used to finance the accumulation of human capital. Education expenditures have to be treated as investment and not as consumption; they durably improve the individuals' productive capacities. In this "augmented" Solow model, the elasticity of per-capita GDP to the savings rate is no longer $\alpha/(1-\alpha)$ but $\alpha/(1-\alpha-\gamma)$ where γ is the share of human capital in the production function. For a value of γ close to 0.5, the model becomes realistic.[28] The model therefore predicts that convergence of per-capita GDP is conditional on the proportion of income invested in physical and human capital. So countries that do not invest in education cannot converge toward developed countries whatever their physical investment. Moreover, Mankiw et al. find that the predictions of the standard Solow model without human capital are plausible (they lead to $\hat{\alpha} = 0.36$) when the sample is limited solely to the OECD countries. This suggests that there is unconditional convergence within the OECD but not in the world economy. The three authors explain it by the fact that the levels of accumulated human capital are comparable among OECD countries.

Yet, the Mankiw et al. model does not provide a fully satisfactory explanation of growth. Like in the Solow–Swan model, of which it is only an extension, growth in the steady state depends on exogenous factors (demography and technical progress) only. However, it shows that a simple extension of the standard Solow–Swan model makes it capable of accounting for the complexity of the catching-up process and of explaining an important part of growth divergences between countries. In practice, the absence of unconditional convergence among all countries in the world illustrated by figure 6.5 would be mainly explained by differences in the rate of accumulation of human capital.

6.2.2 External effects, innovation and growth

The Mankiw et al. approach suggests that in order to better understand the origins of growth, the TFP "black box" must be opened. The corresponding theories were born in the 1980s and 1990s and are known as endogenous growth theories.

There are at least two good reasons to suppose that TFP is not an exogenous phenomenon:

1. Productive efficiency does not rely on the sole efforts of each firm but also on the interaction between them: The accumulation of "know-how" and the benefits from agglomeration, such as the attraction of skills, the development of specialized suppliers, etc. These external effects explain why geographical clusters emerge

28. This is because $0.3/(1-0.3-0.5) \sim 1.5$.

and grow. They need to be incorporated into the theory in order to understand how the organization of markets impacts on growth and when public intervention is necessary.

2. Technical progress results from major inventions and from innovations that naturally depend on the overall scientific context[29]—and perhaps also on luck—but inventions and their application in industry also respond to economic constraints and incentives: Firms invest in R&D to create new products that will give them a competitive edge; as a consequence, the pace of innovation cannot be regarded as given, and it can be built into utility-maximizing economic models.

The common feature of endogenous growth models is to relax the hypothesis of a decreasing return to capital at the aggregate level. Growth can, therefore, be self-sustained, even in the absence of exogenous technical progress.

a) Externalities

External effects are at the root of the first *endogenous growth** models, the first of which was developed by Paul Romer (1986) and is presented in a simplified form in box 6.9. The key idea is that in the presence of external effects, the *social* return to capital is higher than the *private* return because investment has positive effects beyond those the investing company can appropriate. Hence, the return to capital may be decreasing at the firm level, but constant economy-wide.

Telecommunication networks provide a good example of such mechanisms. To each user, connection to a network (either for voice communication or, for example, for exchanging music) gives access to transactions with all other connected users. Such access represents for each individual the *private* profit of being connected. However, the connection of an additional user increases the usefulness of the network for each already connected user.[30] Every additional connection therefore produces a positive externality, which means that the *social return* it generates is thus higher than the private return to the new user. This is known as a *network externality**.

More generally, investment carried out by a specific firm often generates positive spillovers onto other firms. For instance, the investing firm needs to train its employees in the new technologies embodied in the new generation of capital. This "know-how" will later be available to other firms through labor mobility and contacts along the supply chain. This *learning-by-doing** process, already formalized by Kenneth Arrow in 1962, forms the basis for Paul Romer's model presented in box 6.9. Romer's "know-how" resembles Mankiw et al.'s "human capital." A crucial difference, however, is the presence of externalities that allow the economy to escape the curse of a decreasing

29. See, for example, Kuhn (1962).
30. Congestion costs are ignored.

return to capital.[31] As a result, GDP growth can be sustained even in the absence of exogenous TFP growth.

Note that in box 6.9, since each firm remunerates capital at its marginal productivity, the share of capital income in total income is α, like in the Solow model. Network externalities and know-how are not remunerated: These are public goods freely accessible to all. As a consequence, there are no private incentives to develop them and public policies play an important role in allowing them to fully come into play. This will be discussed in section 6.3.

Box 6.9 Learning-by-Doing and Growth in the Romer Model

In Paul Romer's (1986) "learning-by-doing" model, the economy is made up of N identical firms under perfect competition. Each individual firm operates with a Cobb–Douglas production function so that the production of firm i at any time t is written as:

$$Y_{it} = A_t K_{it}^{\alpha} L_{it}^{1-\alpha} \tag{B6.9.1}$$

Total factor productivity A is not exogenous but depends on the economy's total capital stock. Romer considers that the size of the productive sector creates a positive network externality through the exchange of know-how, which he designates under the generic term of *learning-by-doing* and which improves productivity. Accordingly, he uses the following specification for TFP:

$$A_t = AK_t^{\beta} \tag{B6.9.2}$$

where $K_t = \sum_{i=1}^{N} K_{it}$ and, since firms are identical, $K_{it} = K_t/N$ for all i.

In the specific case where $\alpha + \beta = 1$ and if all firms are identical, the model evolves to the AK model briefly introduced in box 6.6 where the marginal productivity of capital is constant: Here we have $Y_{it} = AN^{\beta} K_{it} L_{it}^{1-\alpha}$. Unlike in the Solow–Swan model, growth is self-sustained, even in the absence of exogenous technical progress.

The difference of treatment of human capital in the augmented Solow–Swan model à la Mankiw et al. and in endogenous growth models lends itself to empirical investigations. Does it contribute to growth only transitorily (as the former model would suggest) or permanently (as the latter models would imply)? According to recent OECD work (Arnold et al., 2007), growth in OECD countries seems to support the latter rather than the former. This has significant implications for policy, as it suggests that spending on research and

31. Mankiw et al's model in fact becomes an endogenous growth model when the sum of the shares of both factors—physical capital α and human capital γ in the production function is equal to unity.

education can have a lasting impact on economic growth (rather than on the sole level of per-capita income).

A second type of endogenous growth models, illustrated in box 6.10, considers public infrastructures (or, more generally, public expenditures on education and public services) as an additional production factor able to prevent marginal returns to private capital from falling. Public infrastructure plays the role of the know-how of box 6.9. It is a factor in long-term growth, but through its impact on supply rather than on demand—unlike in the Keynesian models studied in chapter 3. As we shall see in section 6.3, such models provide a rationale for infrastructure policies, public investment in research, and official development assistance to poor countries.

There is a limit, however, to the ability of public investment to support long-term growth. Any public expenditure is financed by a tax on (present or future) privately created wealth; this tax reduces the net return on investment and slows down private capital accumulation. Hence, there is a trade-off between, on the one hand, the provision of productivity-enhancing public infrastructures, and on the other hand, the introduction of a distortion likely to lower production. Box 6.10 illustrates this trade-off in a simple model of growth. At the optimum, a rise in public expenditure increases output by a quantity that is exactly sufficient to finance this additional expenditure.

A number of empirical studies have estimated the impact of the accumulation of public capital on GDP per person and found it to be significant. In the US case, a 10% rise in the stock of public capital was found to translate over the long run into a 4% increase of per-capita GDP (Munnell, 1992). This elasticity, however, is lower, around 0.2 for regions or municipalities.

Box 6.10 Public Intervention and Long-Term Growth

The model, based on Barro and Sala-ì-Martin (1995, ch. 4), highlights the trade-off between positive externalities generated by public expenditures and taxation-induced distortions. It features an economy where public R&D and education expenditures G are financed by a value-added tax τ and raise total factor productivity:

$$G = \tau Y \tag{B6.10.1}$$

$$Y = (AG^{1-\alpha})K^{\alpha} \tag{B6.10.2}$$

where K is the physical capital stock. To simplify, the labor force is supposed constant and equal to unity. From these two equations, the aggregate relationship between K and Y can be written as:

$$Y = \tau^{\frac{1-\alpha}{\alpha}} A^{1/\alpha} K \tag{B6.10.3}$$

Production exhibits constant returns to the physical capital stock: As in box 6.9, it is an AK-type model in which long-term growth is possible even without technical progress.

Equation (B6.10.3) says that, for a given level of private capital K, an increase in the tax rate τ raises output Y. However, K is not constant when τ increases, because a rise in τ reduces the marginal return on capital. Indeed, under profit maximization, K is set at a level that allows the after-tax marginal return on capital to be equal to the cost of capital, i.e., to the sum of the interest rate r and of the depreciation rate δ:

$$r + \delta = (1 - \tau)\frac{\partial Y}{\partial K} = (1 - \tau)^{\frac{1-\alpha}{\alpha}} A^{\frac{1}{\alpha}} \qquad \text{(B6.10.4)}$$

In a closed economy, the interest rate is given by (B6.10.4). It is a hump-shaped function of the tax rate τ: For $\tau > 1 - \alpha$, the interest rate increases with τ; beyond $1 - \alpha$, r decreases when τ rises further. Assuming the savings rate σ to be a monotonic, increasing function of r, capital accumulation is written as:

$$\dot{K} = \sigma(\tau)Y - \delta K \qquad \text{(B6.10.5)}$$

where $\sigma(\tau)$ follows the same hump shape as $r(\tau)$. Assuming a constant value for both A and τ, the growth rate of K and of Y is the same (see equation (B6.10.3)), equal to:

$$g = \frac{\dot{K}}{K} = \sigma(\tau)\tau^{\frac{1-\alpha}{\alpha}} A^{\frac{1}{\alpha}} - \delta \qquad \text{(B6.10.6)}$$

Even without any exogenous technological trend, growth can still be positive in the long run in the presence of public intervention. The relation between the tax rate, τ, and the growth rate, g, is however nonlinear. In Barro and Sala-ì-Martin (1995), the savings rate derives from utility maximization. Growth is maximized at $\tau^* = 1 - \alpha$.

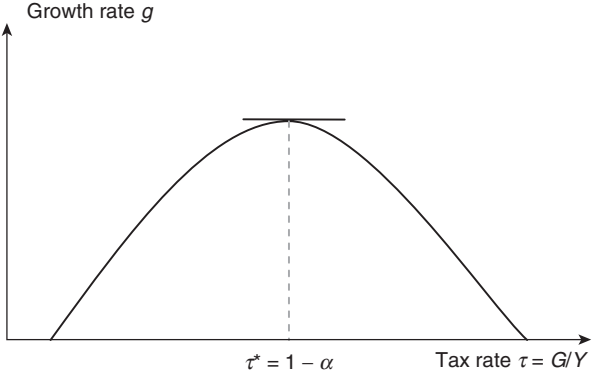


Figure B6.10.1 Taxation and growth.

This first series of endogenous growth models justifies public intervention on two conceptually different grounds: On the one hand, in order to coordinate private decisions so as to exploit externalities among economic agents, and on the other hand, in order to produce public goods—infrastructures, education, public research . . . —which enhance private productivity. A model like the one presented in box 6.10, however, suggests that state intervention can either be favorable or detrimental to growth depending on the level of taxation.

b) Creative destruction

The Austrian-born economist Joseph Schumpeter—a man who had a huge influence on the economics of innovation—identified five types of innovation: (i) On products, (ii) on methods, (iii) on demand, (iv) on raw materials, and finally (v) on firms' organization. In his 1942 book *Capitalism, Socialism and Democracy*, Schumpeter analyzed the process of *creative destruction** through which a major innovation leads to the disappearance of the previous generation of products. Entrepreneurs engage pecuniary and human resources to find and exploit new technologies. They are constantly likely to be dispossessed by competing innovation, but until competing innovation is there, the innovative company remains profitable. The expectation of profit creates an incentive to innovate. Since profit is built on the elimination of the previous generation of innovations, Schumpeter called this process creative destruction.

> These revolutions periodically reshape the existing structure of industry by introducing new methods of production—the mechanized factory, the electrified factory, chemical synthesis and the like; new commodities, such as railroad service, motorcars, electrical appliances; new forms of organization—the merger movement; new sources of supply—La Plata wool, American cotton, Katanga copper; new trade routes and markets to sell in and so on. . . . Thus there are prolonged periods of rising and of falling prices, interest rates, employment and so on, which phenomena constitute parts of the mechanism of this process of recurrent rejuvenation of the productive apparatus.
>
> Now these results each time consist in an avalanche of consumers' goods that permanently deepens and widens the stream of real income although in the first instance they spell disturbance, losses and unemployment. . . . the capitalist process, not by coincidence but by virtue of its mechanism, progressively raises the standard of life of the masses. It does so through a sequence of vicissitudes, the severity of which is proportional to the speed of the advance. But it does so effectively.
>
> Joseph Schumpeter (1942/1976), p. 68

The central role assigned by Schumpeter to the entrepreneur was criticized by French historian Fernand Braudel, who advocated a more systemic

approach prefiguring the importance that economists would give to institutions in the 1990s and 2000s (cf. infra):

> I do not believe that Josef Schumpeter was right to consider the entrepreneur as the deus ex machina. I obstinately believe that it is the overall motion that is crucial and that any capitalism is in the first place the reflection of the underlying economies.
>
> Fernand Braudel (1985), p. 67 (authors' translation)

Creative destruction has major policy consequences. It implies that declining industries should not be protected. On the contrary, the displacement of existing firms and industries by newcomers should be encouraged as an engine of innovation and economic growth. Implementing such a philosophy has proven difficult, however, since it relies on the adjustment mechanism by which redundant employees in declining industries will find jobs in the new industries. In continental Europe, labor mobility (both geographically and between sectors) is limited and labor force reallocations are generally accompanied by substantial wage losses. Moreover, job destruction is immediate while the "creation" is slow to materialize. This makes such adjustment painful and often politically unacceptable.

The recognition of the creative destruction process at the microeconomic level also sheds light on the sources of productivity divergences between countries. For example, research conducted at the OECD (Bartelsman et al., 2003) has highlighted three salient facts:

- In the developed economies, about one-third of labor-productivity gains come from churning, i.e., from the creative destruction of firms (the remaining two-thirds being achieved within existing firms). Firms' demography therefore appears as an important determinant of economic growth.
- New and old firms do not equally contribute to productivity gains. Old firms increase productivity through investing and substituting capital for labor. New firms typically raise TFP. The renewal of firms therefore in itself contributes to productivity gains.
- There is a major difference between Europe and the US. The firms' birth and mortality rates are broadly similar, but surviving firms grow much faster in the US: They are born small but those that survive have more than doubled their labor force over their first two years. In Europe, they grow by 10% to 20% only. In other words, the US economy "tests" new firms and enables them to grow very fast when they introduce innovative products or efficient technologies.

In the mechanism of innovation, competition on the goods market and the protection of intellectual property play a decisive role. Innovation can be seen first as widening the range of products available through so-called *horizontal differentiation**. This mechanism is related to trade liberalization and is

further described in the next section. According to a second approach, very close to the original Schumpeterian vision, innovation consists in improving product quality through what is known as *vertical differentiation**. Every new product moves the technological frontier and eventually completely displaces the previous one while squeezing the rents accruing to their producers and opening new profit prospects to innovators. A recent example is the development of digital photography and the resulting displacement of film photography. In this spirit, a model by Philippe Aghion and Peter Howitt (1992), shows how the R&D effort—and therefore eventually the growth rate of the economy—depends on the expected gains from innovation (box 6.11). Their model predicts that the innovation effort is less when innovation is more easily replicated (in the absence of a patent system) but also when competition on the goods market increases (because the innovation rent decreases). Section 6.3 will further elaborate on these two conclusions and discuss their consequences for public policy.

Box 6.11 The Economics of Innovation in the Aghion–Howitt Model (1992)

The model focuses on the determinants of the research and development effort and on its effects on growth. This box provides a simplified version.[a]

Labor is the only factor of production and can be used either in the production of consumer goods or in research, the latter producing innovations that increase productivity.

The total supply of working hours L is therefore allocated either to production, for a quantity X, or to research, for a quantity N. Hence:

$$X + N = L \qquad \text{(B6.11.1)}$$

Consumer goods are produced by firms under perfect competition according to the following technology:

$$Y = AX^{\alpha} \text{ with } A > 0,\ 0 < \alpha < 1 \qquad \text{(B6.11.2)}$$

where Y is output.

Productivity is represented by the variable A and is endogenous: It can be raised by innovations, which stem from research. However, research results are random: A unit of labor employed in research produces with a probability $\lambda < 1$, an innovation that improves productivity by a factor $\gamma > 1$. The parameter γ therefore measures the size of innovations and λ their frequency.

Labor-market equilibrium requires that the expected return to research equals the hourly real wage w. If $\pi(\gamma)$ represents the expected profit from innovation (λ being the probability to achieve it) we thus have:

$$w = \lambda\pi(\gamma) \qquad \text{(B6.11.3)}$$

If the research effort is successful in producing an innovation, the innovator is then the sole person to command a superior technology. He or she benefits from it by eliminating existing firms and immediately deriving a profit π:

$$\pi(\gamma) = \gamma A X^{\alpha} - wX \tag{B6.11.4}$$

However, this gain is temporary: At the next period, innovation is fully disseminated, and the rent from innovation disappears.

If the research effort was unsuccessful in the first place, the quantity of labor devoted to it brings no return.

Profit maximization leads to:

$$\frac{d\pi}{dX} = \alpha\gamma A X^{\alpha-1} - w = 0 \tag{B6.11.5}$$

Equation (B6.11.5) provides the optimum level of A. Profit then is written as:

$$\pi = \frac{1-\alpha}{\alpha} wX \tag{B6.11.6}$$

Combined with (B6.11.3) this equation leads to the optimum allocation of labor:

$$X = \frac{1}{\lambda}\frac{\alpha}{1-\alpha} \quad \text{and} \quad N = L - \frac{1}{\lambda}\frac{\alpha}{1-\alpha} \tag{B6.11.7}$$

The amount of labor allocated to research logically depends positively on the probability of success λ (it does not depend, however, on the size of innovations γ since, in equilibrium, productivity earnings are passed to employees; the innovator's profit only comes from his or her displacing the existing producers and appropriating their profits).

In this simple economy without demography or capital, the rate of growth of output is simply the growth rate of productivity resulting from the innovation process:

$$g = \lambda N(\gamma - 1) = \left(\lambda L - \frac{\alpha}{1-\alpha}\right)(\gamma - 1) \tag{B6.11.8}$$

The growth rate eventually depends on the probability and on the size of innovations, as well as on the size of the economy (the larger the economy, the higher the return to innovation) and on the share of profits in value added (a higher share encourages innovation because the corresponding rent is larger). The model can easily be extended to a situation where the innovator captures the rent only partially, instead of totally displacing existing producers.

[a]The authors are grateful to Philippe Aghion for having shared this simplified version with them.

There is an important theoretical literature on intellectual property, which underlines the difficult trade-off between patent protection (to encourage innovation) and the dissemination of innovation (to promote its adoption throughout the economy). *Ex ante*, policymakers are tempted to promise rents to innovators, *ex post* they are tempted to expropriate them. This problem illustrates the time inconsistency dilemma analyzed in chapter 2.[32]

Finally, innovation and growth models describe the incentives to innovate, but they ignore the way in which innovations are received and disseminated. However, the dissemination of an innovation requires a "critical mass" of users. Paul David, the historian of innovation, explains how the QWERTY keyboard became a standard on American typewriters (David, 1985). When adopted in the 1870s by one of the first typewriter manufacturers, Remington, this keyboard minimized the risk of keys overlapping each other when the user had to type fast. All competitors eventually adopted it. Yet, studies showed that the *Dvorak Simplified Keyboard* (DSK), a system patented in 1932, allowed much faster typing. However, despite its superiority, DSK was not able to prevail. The QWERTY system was extended to computer keyboards even though the initial reason (the overlapping of keys) had long been inapplicable. Paul David uses this example to stress the importance of history in economic choices and the fact that the actual destination often depends on the trajectory (this is called *path dependency**), while economists too often describe equilibrium situations without taking the initial situation nor the trajectory into account. Path dependency generally characterizes any innovation that involves network externalities.

6.2.3 Beyond the production function

a) International trade

For a long time, growth theory and trade theory have developed as two separate branches. Growth models were initially developed in a closed-economy framework, and trade models hardly addressed growth.[33] It is only recently that models have been developed that allow understanding of the relationship between growth and trade.

Beyond the traditional efficiency gains from trade due to specialization, captured in the classical trade models, the relationship between trade and economic growth can be analyzed along three main dimensions. First, there are productivity gains to be expected from heightened competition through trade liberalization. Not only does competition increase the pressure for firms to innovate in order to stay ahead of new foreign competitors, but

32. See for example Guellec (1999) or Tirole (2003).

33. An exception was the "*immiserizing growth*"* model introduced by Jagdish Bhagwati in the late 1950s. Bhagwati, a trade economist, pointed out that growth in a country's export supply could result in a deterioration of the relative price of those exports and that this terms-of-trade effect could result in a deterioration of income.

it also sustains a Darwinian process through which only the fittest, i.e., the most productive firms survive and expand. Second, international trade fosters knowledge spillovers that enhance productivity in the less-advanced countries and sectors. Third, international trade increases the size of markets, which both allows domestic firms to exploit economies of scale (notably through learning by doing), and increases the potential rent accruing to successful innovators (see the model in box 6.11).

The influence of international trade on product innovation is readily understood in the framework of models of trade in varieties of similar products. In those models (introduced in the 1980s), consumers choose between products (say, cars or restaurant meals) according to their preferences and relative prices but they also choose between varieties of the same products (say, Toyotas or Volkswagens and sushis or sashimis). The larger the range of varieties available, the greater the consumers' utility: Consumers are said to have *taste for variety**.

Love for variety can result from an exogenous preference of the consumer for a diversified consumption basket (for example, as regards food or cultural products), or from a trial-and-error research into the ideal variety (for example, for the purchase of a car).

Formally, the consumers' utility is often assumed to be represented by a *Dixit–Stiglitz**[34] function, which makes utility dependent on both the overall quantity consumed and the number of products available to consumers. Assuming there is a continuum of goods indexed on [0, 1], and calling C_i the consumption of good i, consumer utility $U(C)$ is written as:

$$U(C) = \left[\int_{i=0}^{1} \alpha_i C_i^{(\sigma-1)/\sigma} \, di \right]^{\sigma(\sigma-1)} \quad \text{where} \int_{i=0}^{1} \alpha_i di = 1 \qquad (6.7)$$

where σ represents the elasticity of substitution between products and α_i the weight of good i in the consumer's utility.

Innovation can be regarded as consisting in widening the range of varieties available to consumers. The food industry provides a good example of this sort, since a large part of innovation in this sector consists in simply extending the variety of goods available to consumers (new yoghurt flavors or textures, for example).

In a closed economy, the expansion of varieties is bound by a trade-off between efficiency in production and the number of varieties produced. A simple way to represent this is to assume that the production of each variety involves a fixed as well as a variable cost. Producing more varieties is then detrimental to productivity.

34. Named after the seminal contribution by Avinash Dixit and Joseph Stiglitz (1977) which expands on a monopolistic competition framework initiated in 1933 by Chamberlin. See Krugman (1995) and Combes et al. (2006) for a history of these ideas.

Trade, however, allows the specialization of producers and countries and the exploitation of economies of scale through access to larger markets. Under free trade, each country produces fewer varieties but consumers have access to more. International trade allows reaping the benefits from economies of scale in the production process without restraining consumer choice.

Now, let us assume that the producer of each variety enjoys some monopoly power because his or her output is not perfectly substitutable for other products. This provides both an incentive for product innovation, which is then guided by the quest for the rents generated from such monopoly power, and a mechanism for endogenous growth, since the return to capital is now increasing. The model thus describes a self-sustained growth process driven by the specialization of the labor force in a constantly increasing range of goods exhibiting increasing returns. The intuition for the mechanism goes back to an article by Alwyn Young in 1928; Romer (1990) and Grossman and Helpman (1989) have provided examples of such models.[35]

This approach highlights the importance of demand externalities (as opposed to the previously described production externalities): A wider market generates a larger solvent demand for each product variety, which stimulates output and distributed income. The existence of this "virtuous circle" also sheds some light on the reasons why some countries may remain trapped in underdevelopment: Their domestic market is simply too small to generate the necessary investments.

b) Geography and history

Growth theory studies the evolution of wealth over time. Its distribution across space was long ignored by classical economists but, following Hotelling's (1929) seminal work on spatial competition, it has elicited a growing interest since the 1950s. Starting in the 1990s, research has explored the interactions between growth and geography.[36]

The supply and demand externalities highlighted by endogenous growth models explain why growth rates differ across countries and regions: Firms choose their location according to geographical (access to transportation infrastructures, to natural resources, to drinking water, etc.), cultural (language, political system), and industrial (proximity to suppliers, access to final consumers, know-how externalities) criteria. Understanding such mechanisms has been the focus of the *new economic geography** after Paul Krugman (1991a, 1991c) outlined this research program in the early 1990s.

The toolkit of the new economic geography resembles that of endogenous growth theory: The assumptions of monopolistic competition and/or externalities open the way to increasing returns and to the notion that a "critical mass" of activities and product differentiation gives an important

35. See Gancia and Zilibotti (2004) for a detailed review.
36. See Combes et al. (2006) for a history of the relations between economics and geography.

role to the size of markets. The specific feature of economic geography is the introduction of transport and congestion costs that may offset the incentives for concentration. Firms face a trade-off between concentrating their activities in a single location to take advantage of economies of scale, and disseminating them to reduce transport costs and get closer to end-consumers. Overall, the spatial location of economic activities results from a balance between forces of agglomeration and forces of dispersion.[37]

As a consequence, there is no single growth trajectory, unlike in the Solow–Swan model, but growth paths exhibit multiple equilibriums and path dependence: *History matters.* The *core/periphery model**, revisited by Krugman (1991b) and described in box 6.12, provides an example of such thinking. The model formalizes the tension between forces of dispersion resulting from transport costs and forces of agglomeration resulting from access to larger markets and the existence of fixed costs in the production process. When transport costs are high, the former dominate and production remains scattered. When transport costs are weak, the latter prevail and production concentrates in one place. The equilibrium is discontinuous, even *catastrophic* in mathematical terms:[38] A small change in transport costs can lead to a brutal relocation of firms.

These mechanisms shed light on past economic history. Why are Hong Kong and Singapore major financial centers? Because both cities developed around their harbor and warehouses and had both a "critical mass" of capital and skilled labor accumulated throughout the twentieth century, which they could shift toward new activities in the 1970s and the 1980s. Why could peripheral countries like Finland and New Zealand develop in the late twentieth century? The answer is that transport and communication costs decreased dramatically. Why is sub-Saharan Africa so poor? In large part because of the legacy of colonialism, and because many African countries are landlocked and too far from dominant markets to be competitive in spite of abundant natural resources and low labor costs. Looking forward, if commodity prices become higher as resources are depleted, the associated rise of transport costs may weaken agglomeration forces in the world economy.

Economic geography also has prescriptive implications. The public sector can influence firms' location decisions; this is why European governments compete to attract company headquarters and regularly quarrel over the location of regulatory authorities: Every country hopes that by so doing, it will increase its attractiveness.

37. The handbook by Baldwin et al. (2003) presents these models. Krugman (1995) reviews the history of the theory of development in the light of these mechanisms.

38. A catastrophe, or bifurcation, is a noncontinuous jump from one path to another in a nonlinear dynamic model with several possible paths.

Box 6.12 The Core/Periphery Model

The Core/Periphery (CP) model[a] describes two identical regions ("North" and "South"), each having two sectors (agriculture and manufacturing). Production in agriculture exhibits constant returns to scale under perfect competition. The manufacturing sector is characterized by product differentiation (there are n varieties) under monopolistic competition.

Production involves a fixed cost, which generates economies of scale, and a variable cost which is a function of the production level. Both goods are traded between the two regions. There is free trade in agriculture, but trade in manufactures involves trading costs, so that that τ units of manufactures must be transported for one unit to arrive at destination.[b] The degree of openness ϕ is therefore a decreasing function of the transport costs τ.

Consumers are assumed to exhibit a preference for variety and their behavior is represented by a constant elasticity of substitution utility function à la Dixit–Stiglitz. Utility therefore depends positively on the number of varieties of the goods that are consumed, and the consumer chooses his or her consumption basket in two stages: S/he first allocates income between the agricultural and manufactures, and then chooses the quantity and number of varieties of manufactures. Agricultural labor is immobile while industrial workers can migrate from one region to the other.

The dynamics of location results from the combination of three effects:

- A "market-access" effect: Firms tend to locate in the big market and export to small markets (thus saving on transport costs).
- A "cost-of-living" effect which reflects the impact of firms' locations on the local cost of living. For example, spatial concentration in the North leads to higher real income in that region because northern consumers import less and save on transport costs.
- A "market-crowding" or congestion effect: Imperfectly competitive firms tend to locate in regions with fewer competitors, and competition between firms leads to higher wages and makes agglomeration less attractive.

If workers are mobile, the first two effects can mutually reinforce themselves in a "cumulative causality" dynamics leading to spatial concentration. For example, if there is a shock that leads to migration from South to North, the market-access effect will encourage some manufacturing firms to relocate in the North; the cost-of-living effect will in turn mean that a given nominal wage will have more buying power in the North, thus inviting further migration. These two effects represent agglomeration forces. Conversely, the market-crowding effect acts as a dispersion force.

The actual location of manufacturing thus depends on the relative strength of the agglomeration versus dispersion forces. If the former prevail, any migration shock will result in all industrial workers and firms moving to one region. Conversely, if the latter prevails, the symmetric equilibrium is stable: Any movement of labor will be offset by a reverse movement due to the market-crowding effect.

Baldwin et al. (2003) show that trade costs affect both agglomeration and dispersion: The freer the trade, the smaller the forces both of dispersion and of agglomeration (since the distinction between the two regions tends to disappear). The detailed model shows, however, that dispersion dominates when trade costs are very high, while a reduction in trade costs weakens dispersion more rapidly than agglomeration. As trade costs decline, they reach a level at which dispersion and agglomeration forces balance each other. More precisely, there are two important thresholds for the degree of openness ϕ (or equivalently for trade costs as ϕ is a function of τ:

- The *break point* ϕ^B is the threshold beyond which the core-periphery outcome where all manufacturing production is located in one region is stable. Beyond ϕ^B, agglomeration dominates dispersion so that a symmetric distribution of production is impossible because any shock gives rise to all manufacturing moving to one region; below ϕ^B, the symmetric equilibrium is stable.
- The *sustain point* $\phi^s < \phi^B$ is the threshold below which only the symmetric equilibrium can be observed. Neither of the two agglomeration equilibriums is possible because dispersion forces are dominant.

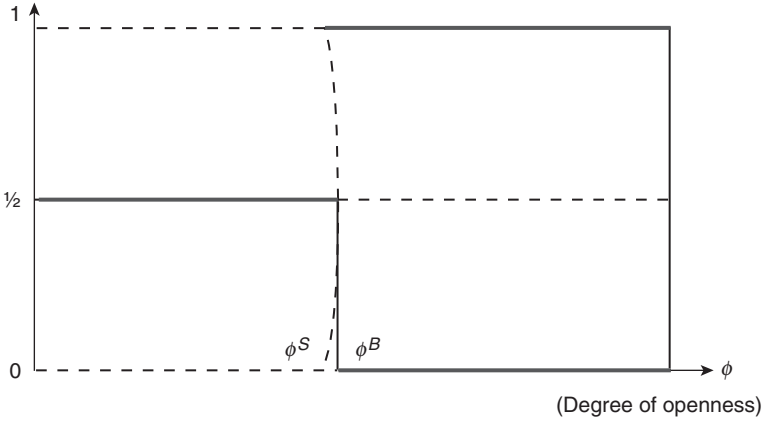


Figure B6.12.1 The *Tomahawk* diagram.
Source: Baldwin et al. (2003).

Between ϕ^s and ϕ^B, all three outcomes (the two core-periphery equilibriums and the symmetric equilibrium) are possible and stable. Agglomeration may occur before the degree of openness reaches the break point, when openness is not high enough for the symmetric equilibrium to be unsustainable and not low enough for agglomeration to be dominated by dispersion. Once workers and industries are agglomerated, they have no incentive to move, because the marginal gains from higher relative wages would be offset by transport costs.

One of the interesting features of this model is that a parallel increase in openness of initially symmetric regions leads to asymmetry. The three possible outcomes are represented by the full lines in figure B6.12.1 (the so-called *Tomahawk* diagram), where the X-axis represents the degree of openness ϕ, and the Y-axis the fraction (ranging between 0 and 1) of the manufacturing workers located in the North. In the areas where several equilibriums are possible, for a degree of openness beyond the sustain point and below the break point, a temporary shock, or even a simple change of expectations, can result in shifting from one equilibrium to another. Migration and agglomeration display a catastrophic nature.

Although it is highly stylized, the model captures one of the deep insights of the new economic geography: Location is determined by both deterministic and random factors. The reason why the US movie industry is based in Hollywood is that it migrated there from New York in the 1910s after Thomas Edison and a few other companies had liaised to exploit their technological monopoly and had established a centralized patenting system. Independent companies unwilling to abide by the rules set by what had become known as *The Trust* migrated west and soon settled in Hollywood where D.W. Griffith had shot a movie in 1910. Los Angeles's climate was certainly a factor in the choice of this location but many other places could have been chosen. However, after the agglomeration effects had been set in motion, the industry quickly concentrated there and the dissolution of the patent oligopoly in 1918 did not reverse the trend. Likewise, all cities were initially founded on the basis of such geographical criteria as access to rivers or elevated watch points, which have become less relevant over time.

[a] Our presentation is drawn from Baldwin et al. (2003, ch. 2).

[b] This is called the *iceberg* model of trade costs because part of the product "melts" during transport.

c) Income distribution

The relationship between income distribution and development has been the subject of intense debate.[39] We documented at the beginning of this chapter how growth was accompanied by an increase in world inequality during the nineteenth and twentieth centuries. Conversely, inequalities affect growth through several channels that economic theory has attempted to clarify.

There are many reasons why inequality may adversely impact growth:

- Income inequality often translates into an inequality of opportunities. In particular, less-developed countries have underdeveloped financial markets. Exclusion from credit markets prevents the poorest individuals from investing, whether in physical or in human capital (education), which in turns locks them in *poverty traps**—hence the interest in micro-credit pioneered by Muhammad Yunus as a way to relax the credit constraint on the poor.[40]
- Income inequalities may lead to political instability or political deadlock. The risk of misery-based riots or revolutions creates a climate of uncertainty that discourages investment.
- In a democracy, inequality may tilt the political balance toward redistribution rather than toward incentives to wealth creation. For example, Benabou (1996) presents a theoretical model in which income dispersion increases the risk of conflict between social groups over the distribution of profits, and creates a "prisoner's dilemma" in which none of these groups wishes to contribute to wealth creation. Alesina and Rodrik (1994) emphasize another mechanism based on tax incentives: The more uneven the primary distribution of income, the more the median voter will vote for a redistributive taxation. However, an excessively high marginal tax rate on high incomes is a barrier to capital accumulation and therefore to growth.

Conversely, in the absence of redistributive taxation, an increase in inequality can be favorable to growth if wealth accumulated by the richest fraction of the population is invested in the industries that generate productivity gains. In turn, those gains may "trickle down" to the less wealthy. This story is consistent with the Kuznets curve introduced in the first section of this chapter, but there is little empirical support for any automatic trickle-down mechanism.

In many countries, those issues are a matter for fierce policy debates. A tentative conclusion based on the available empirical evidence is that inequalities may have a negative influence on growth in underdeveloped economies, but a positive one in developed countries. Having built a very rich dataset, Deininger and Squire (1996) conclude that the relation between

39. See the Kanbur (2000) synthesis.

40. Rajan and Zingales (2003) highlight that the lack of access to finance is a key determinant of the persistence of poverty.

inequality (measured by the Gini index of income distribution) and the growth rate depends on the development level: They find a negative influence of inequalities on growth for either low or high GDP per capita, and a positive influence in between.[41] However, they find that the inequalities that hamper growth are not income inequalities but rather factor endowment inequality, especially as regards land distribution.

d) Institutions

So far, we have primarily associated TFP growth with technical progress. However, TFP depends, in a much more general way, on all factors that contribute to raising the effectiveness of labor, capital, and their combination. Important dimensions here are the legal and regulatory environment of production, the nature of the relationship between employers and employees, the enforceability of laws and contracts, all factors that can be summarized under the generic term of *institutions**. Douglass North, who was awarded the Nobel Prize in 1993 with Robert Fogel, has defined the institutions as "the humanly devised constraints that structure human interaction. They are made up of formal constraints (rules, laws, constitutions), informal constraints (norms of behavior, conventions, and self-imposed codes of conduct), and their enforcement characteristics" (North, 1990). Following Ronald Coase, the emphasis here is on the transaction costs implied by a low-quality institutional environment and on the importance of the security of contracts.[42] Thus, the more uncertain the legal, tax, and social environments are, the larger the precautions that any given investment requires.

In an influential contribution, La Porta et al. (1999) have stressed the importance of *legal origins**. In their view, countries such as France and the former French colonies with a civil law tradition suffer from an overextended government and regulations hampering private initiative, while the UK and its colonies operating under common law benefit from more flexible institutions and a better protection of property rights. According to the authors, such difference can be traced back to the different contexts of France and England in the twelfth and thirteenth centuries, the former being decentralized and prone to rebellion while the latter was calmer and more industrious. La Porta et al. (1998) have also suggested that civil law is more likely to be associated with intermediated finance, while common law better underpins market finance since it better protects minority shareholders

One can object that countries like China and India have developed original models which cannot be reduced to civil law or common law, and that there is always a gap between formal legal principles and on-the-ground experience. As Dani Rodrik (2004) has suggested in the case of China and Russia:

41. Also see Banerjee and Duflo (2003) for a discussion of the methods used.
42. See North (1990) and the literature review in Borner et al. (2003).

> In Russia, an investor has in principle the full protection of a private property-rights regime enforced by an independent judiciary. In China, there is no such protection, since private property has not been (until very recently) legally recognized and the court system is certainly not independent. Yet during the mid- to late-1990s, investors consistently gave China higher marks on the rule of law than they did Russia. They evidently felt better protected in China than they did in Russia.
>
> Dani Rodrik (2004)

Daran Acemoglu, Philippe Aghion and Fabrizio Zilibotti (2002) have proposed an interesting framework of analysis by introducing the concept of *distance to the frontier**: For countries farthest away from the technological frontier, technical progress mainly takes place through the adoption of existing technologies, and the institutions favorable to growth are those that encourage this imitation process; but the closer one gets to the frontier, the more important it is to encourage innovation and to develop specific institutions capable of protecting intellectual property, fostering project finance, or giving incentive to risk-taking.

This analytical framework can easily be transposed to other fields. It came to be understood in the aftermath of the 1997–98 financial crises in emerging market economies that the opening of the financial account should not be recommended to all countries, as the OECD and the IMF tended to believe before the crises, but only to countries equipped with robust financial institutions (Kose et al., 2006). The main contribution of Acemoglu et al. (2002) is finally to show the importance of flexible institutions. Institutions matter at all stages of development but they must adapt to each stage. This is an invitation for international institutions to refine their recommendations to developing countries.[43] Together with J. Wallis and B. Weingast, D. North has developed a theory of development as a transition process between institutions (box 6.13).

Box 6.13 Douglas North's Approach to the Social Development Process

According to North (1994, pp. 4–5), "it is adaptive rather than allocative efficiency which should be the guide to policy. Allocative efficiency is a static concept with a given set of institutions; the key to continuing good economic performance is a flexible institutional matrix that will adjust in the context of evolving technological and demographic changes as well as shocks to the system [. . .] It is doubtful if the policies that will produce allocative efficiency are always the proper medicine for ailing economies."

43. See Acemoglu et al. (2004) for a general synthesis on the role of institutions in growth.

In a recent work, North et al. (2006) emphasize the necessary connection between economics and politics within a "social order." The authors consider that only three generic social orders have existed in history:

- The *primitive social order* that dominated pre-recorded human history.
- The *limited access social order* in which violence is contained, and order and stability maintained, through political manipulation based on rent generation through limited entry and rent distribution. This order rests on the logic of the "natural State." In response to endemic violence, warlords agree on controlling and sharing property rights and rents, which creates a common interest in pacifying relations. Access to all functions is limited, and constitutes privileges for those who receive them and share an interest in stability. The limited access order is based on cronyism, personalization and corruption, but the "natural State" is neither fragile nor failing. It simply corresponds to the first stages of social development of societies prone to natural violence. As such, the limited access social order is stable.
- The *open access social order* that emerged over the last 300 years and was adopted by the few countries that successfully developed. It rests on political and economic competition and on the contestability of rents. Rents do exist, but they result from dynamism and innovation, are fundamentally impersonal (rather than attached to a person) and can be contested. They cannot be appropriated forever, and their contestability (through elections or through competition) makes their distribution at any point in time acceptable for all, including those who do not benefit from them. Organizations rest on membership and contract, and their credibility stems from their largely perpetual and impersonal character. The open access social order is also stable.

For North et al., development really means the transition from a limited access to an open access social order. Preconditions for this transition are the emergence of a legal framework that protects the elites' rights (and that can, over time, expand beyond the elites); the emergence of impersonal organizations able to survive individuals; and the political control of the military. In order for the transition to take place successfully, changes must be small, mutually reinforcing, and cumulative. They must also be supported by the ruling elites, so they need to be compatible with the elites' perceived interests, even though the final outcome might not be supported by the elites.

This taxonomy may be oversimplifying (developed countries retain many features of limited access social orders, such as the reproduction of social elites and resistance to the elimination of rents) but it has important implications for policy reform in developing countries. For example, attempts to introduce elements of an open access social order into limited access order societies are bound to fail if the necessary coherence between economics and politics is ignored. Also, a limited access political system is incompatible with economic deregulation and liberalization, and it makes no sense to try to reform it using economic means only.

Empirical studies have endeavored to build indicators of the quality of institutions and to relate them to GDP per person. The World Bank publishes a Worldwide Governance Indicators database with five variables: Voice and accountability, political stability and absence of violence, government effectiveness, regulatory quality, rule of law and control of corruption (Kaufman et al. 2008). Such variables are based on an array of rankings and surveys, and their reliability is therefore debated.

Building on these indicators, the International Monetary Fund (2003) has uncovered a strong positive correlation between the quality of the institutions and GDP per person—which in turn suggests that institutional improvements can foster growth. Figure 6.8 illustrates the correlation across countries between the level of GDP per capita, the quality of regulation, and the respect for the rule of law.

Correlation does not imply causality. Are institutions causing development or vice-versa? It is admittedly easier to convince people to shed their informal protections and networks and trust the rule of law when they can rely on a tax-financed social safety net. This suggests that it may be difficult to find out whether good institutions are conducive to growth, or the other way around. Econometric techniques can be used to sort this out (box 6.14), but only up to a point. The Pandora's Box of the origins of institutions will not be closed anytime soon. There is also another dimension to the debate, namely the nature of the dependent variable: Is institutional quality correlated with the level of income or with the process of economic growth? Meisel and Ould Aoudia (2008) claim that the quality of institutions as measured through the World Bank Worldwide Governance Indicators is strongly correlated with the level of income, but not with the speed of development over a medium-to-long term horizon. They discuss the specific institutional variables that facilitate economic take-off, and those help sustain economic growth over the long term and make economic catch-up possible.

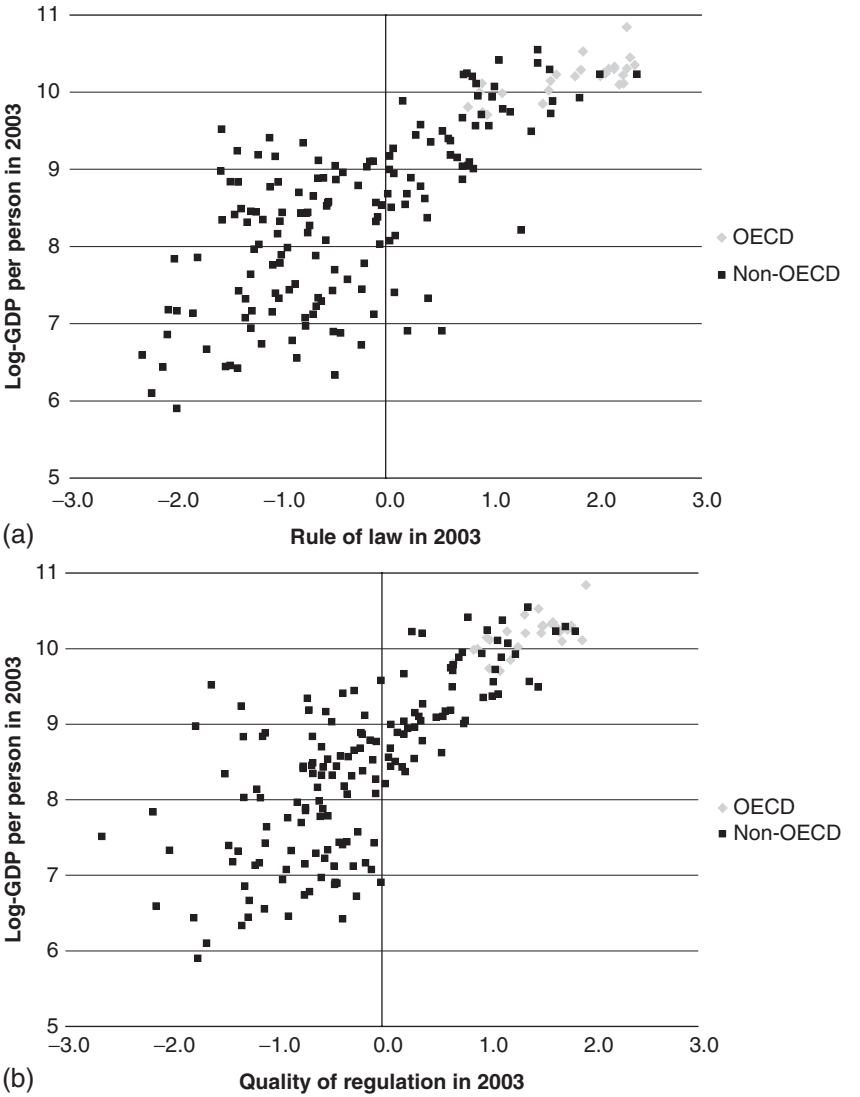


Figure 6.8 Institutions and GDP per capita in 2003. a) Rule of law, b) quality of regulation.
Source: World Bank, Worldwide Governance Indicators, and Heston, A., R. Summers, and B. Aten, Penn World Table Version 6.2, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania, September 2006.

Box 6.14 Do Institutions Cause Growth, or Is It the Other Way Round? Using Instrumental Variables to Identify Causality

Economists tend to think that good institutions help countries grow faster. The problem is that good institutions are easier to set up and sustain in wealthier countries. To sort out this chicken-and-egg problem, an econometrician can rely on information provided by so-called *instrumental variables**, which are variables that are correlated with the quality of institutions (the explanatory variable in our econometrician's equation) but not with the development level (the dependent variable in the equation).

Consider the following regression:

$$y_i = x_i\beta + \varepsilon_i \tag{B6.14.1}$$

y_i is the dependent variable (e.g., GDP per person), x_i is a vector of k explanatory variables (e.g., measures of the rule of law, lack of corruption, degree of democracy, etc.) each being observed at time i, and ε_i is an error term which we suppose to be uncorrelated over time and identically distributed. We intend to estimate β, a vector of k parameters, based on past observations of y and x.

Potential problems with ordinary least squares (OLS) (for a definition of OLS, see box 2.1) are best understood when equation (B6.14.1) is put in matrix form. Let X and Y be the (N, k) and $(N, 1)$ matrices of observed variables, obtained by stacking N observations. Equation (B6.14.1) can be rewritten as:

$$Y = X\beta + \varepsilon \tag{B6.14.2}$$

Multiplying this equation by X', the transposal of X, gives:

$$X'Y = X'X\beta + X'\varepsilon \tag{B6.14.3}$$

The OLS estimate of β is obtained by assuming $X'\varepsilon = 0$ in this equation and by solving the resulting linear system:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y \tag{B6.14.4}$$

We can therefore write:

$$X'X(\hat{\beta}_{OLS} - \beta) = X'\varepsilon \tag{B6.14.5}$$

If x is not correlated with ε, then $X'\varepsilon$ converges toward zero in large samples and $\hat{\beta}_{OLS}$ therefore converges toward β. But if x is correlated with ε, then $\hat{\beta}_{OLS}$ is biased, meaning that it does not converge toward the true value β. The problem is that this often happens when there is uncertainty about the causation behind a correlation. In such cases, relying on OLS estimates can therefore be severely misleading.

An *instrument** is a variable z which is correlated with x (with a correlation matrix of rank k between the variables in z and the variables in x) but uncorrelated with ε. Let R be a (j, k) weighting matrix which we use to select, and possibly combine, the instruments. The instrumental variable (IV) estimator is:

$$\hat{\beta}_{IV} = (R'Z'X)^{-1}R'Z'Y \qquad \text{(B6.14.6)}$$

where Z is the matrix of instruments. Equation (B6.14.5) is then replaced by:

$$R'Z'X(\hat{\beta}_{IV} - \beta) = R'Z'\varepsilon \qquad \text{(B6.14.7)}$$

The IV estimator converges toward the true value β if $R'Z'\varepsilon$ converges toward zero, which is the case since the instruments are not correlated with the error term.

The simplest way to implement IV is called "two-stage least-squares": It involves, first, regressing each endogenous explanatory variable on the full set of exogenous variables (exogenous explanatory variables plus instruments) and, then, estimating the equation, replacing each endogenous explanatory variable by its approximation yielded by the first stage—see chapter 12 of the handbook by Greene (2008).

The search for appropriate instruments involves judgment, as there is a need to find variables which by construction are uncorrelated with the dependent variable.

What does all this tell us about growth and institutions? Daron Acemoglu, Simon Johnson and James Robinson (2001) have estimated the link between the quality of institutions x and today's GDP per person y, using as an instrument z, the prevalence of tropical diseases among settlers in colonial times. Since the local population is immunized against local diseases, the instrument does not reflect local health conditions and is not correlated with subsequent GDP growth, while the authors claim that it has discouraged building durable institutions (through settlements), and is therefore correlated with x. They show that the correlation between institutions and growth is not due to reverse causality. However, as noted by Rodrik (2004), this explanation does not account for the equally successful or unsuccessful growth performances of the many developing countries that were never colonized.

The emphasis on the role of institutions is both an opportunity and a danger for growth theory. The opportunity is to reach a deeper understanding of the determinants of economic performance and to recognize that there cannot be a single institutional template for all countries and at all times. This makes room for much richer policy conclusions. However, if mechanisms are excessively context-dependent, there is a risk of ending up with "soft"

theories which produce neither general testable propositions nor clear policy recommendations. Minimal structure must therefore be imposed on the theoretical description of the link between institutions, organizations, and growth. North's research suggests that this link is complex and nonlinear. This message is increasingly being heard by policymakers, as evidenced by the Spence Report on growth and development commissioned by the World Bank (Commission on Growth and Development, 2008). Section 6.3 explores the resulting policy recommendations.

6.3 Policies

Unlike, say, price stability, for which most countries rely on a single instrument (monetary policy), growth is sought through playing on several keyboards at the same time.

A few months after taking office, French president Nicolas Sarkozy commissioned a report to identify obstacles to economic growth and measures required to lift them. The report came up with 316 recommendations, all deemed indispensable (Attali, 2008), ranging from the competences of local governments (decision No. 260) to the diplomas of hairdressers (decision No. 209). Not all growth strategies look like endless laundry lists, but most are typically wide-ranging and involve the risk of listing individually desirable, but unrelated, reforms. This is particularly evident in Europe. In view of the EU's poor performance in recent decades, growth is a major European priority. In March 2000, European Heads of State and Government met in Lisbon and outlined a growth program with the goal of redressing the EU performance by the year 2010. What became known as the *Lisbon agenda** (box 6.15) is an example of a comprehensive growth strategy that did not deliver on its promises.

Box 6.15 The Lisbon Strategy

In March 2000, the European heads of state and government agreed on "a new strategic goal for the next decade: To become the most competitive and dynamic knowledge-based economy in the world capable of sustainable economic growth with more and better jobs and greater social cohesion."

The Lisbon strategy aimed at:

> "Preparing the transition to a knowledge-based economy and society by better policies for the information society and R&D, as well as by stepping up the process of structural reform for competitiveness and innovation and by completing the internal market;"
>
> "Modernizing the European social model, investing in people and combating social exclusion;"

> "Sustaining the healthy economic outlook and favorable growth prospects by applying an appropriate macro-economic policy mix."

It was added that

> "if the measures [set out] are implemented against a sound macro-economic background, an average economic growth rate of around 3% should be a realistic prospect for the coming years."
>
> Source: Lisbon European Council, 23 and 24 March 2000, Presidency Conclusions, available on the Web site of the European Union.

Five years later, in March 2005, the then enlarged EU observed that "results [were] mixed" and adopted a revised strategy with a stronger focus on growth and employment and a simplified governance. The main objectives, however, were maintained.

The original Lisbon agenda was strikingly ambitious. The "laundry list" syndrome actually materialized, according to an assessment prepared by former Dutch Prime Minister Wim Kok (High Level Group, 2004), which led in 2005 to a refocusing on growth and employment. Yet the revised Lisbon agenda remained ambitious, especially as it requires policy action in fields such as labor markets and research that are not the responsibility of the EU as a whole but of individual governments.

As a technocratic rather than political institution, the OECD may be immune from grand promises. However, in recent years its annual *Going for Growth* report has addressed education, labor markets, pensions, product market regulation, competition policy, and international trade. This is, in fact, hardly surprising in view of the determinants of growth identified in the previous section.

The overriding problem, for any government that wishes to promote growth, is not to find out what needs to be fixed. It is to select priorities. The economist's role is to make the best use of theory to help it.

6.3.1 A roadmap

To put some sort of order in the discussion, we can start from the theories introduced in section 6.2 and sort out policies accordingly (figure 6.9):

- Governments can stimulate labor supply through policies that favor participation in the labor force. Corresponding measures can be regulations (as regards, for example, the retirement age) and changes to tax and benefits rules (as, for example, with the introduction of in-work benefits). In the medium run also, governments can stimulate capital accumulation through tax incentives, competition, and reforms of

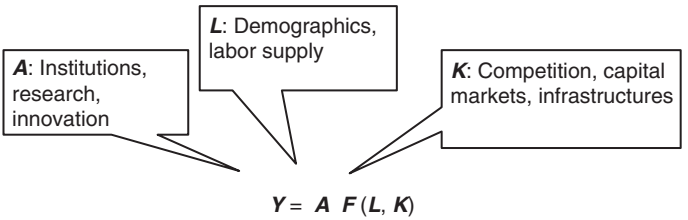


Figure 6.9 Using theory to design growth policies.

financial markets. They can also invest in public capital. The time horizon here is a few years.

- In the long run, (up to a few decades), the capital stock is endogenous and only total factor productivity and labor supply matter. Public policies affect the quality of the labor force through education and training; they also have a bearing on total factor productivity through the funding of research and improvements in institutions.

The public discussion is less clear-cut, as it often confuses long-term and short-term determinants of growth. For example, politicians and voters often attribute long-term economic performance to monetary and fiscal policies. Technocrats tend to hold the opposite view and maintain that macroeconomic policies have no bearing on long-term growth. Both views are equally untrue. Before addressing the levers of a growth program one by one, we first need to clarify the link between short-run and long-run policies.

In the short run (up to a few quarters), supply-side policies are dominated by cyclical fluctuations and by the impact of stabilization policies. In the medium term, however, economists usually assume a clear separation between stabilization and allocation policies (chapter 1). The former are tasked with maintaining production close to its potential level, while the latter aim at raising this potential level. This is, among others, the underpinning of Europe's economic policy framework described in chapter 2. This may not be entirely correct and several arguments point to the existence of an interrelation between long-term trends and short-term fluctuations. These are:

- *Precautionary behavior.* As seen in chapter 4, excessive inflation is bad for long-term growth. More generally, macroeconomic instability leads companies and households to engage in precautionary behavior.[44] We saw also in chapter 2 that uncertainty about the return on investment projects raises their break-even return and delays their implementation. Similarly, increased uncertainty over household income makes individuals consume less and invest more in risk-free (hence unproductive) securities such as Treasury bonds. However,

44. Aversion to risk and precautionary behaviors are explained in chapter 2.

negative-growth consequences of macroeconomic instability may only materialize at high levels of uncertainty—say, two-digit inflation.

- *Unemployment hysteresis.*[45] When employees having lost their jobs in an economic downturn remain unemployed, their skills deteriorate and they become less employable. As time goes by, finding a job becomes increasingly difficult and sometimes even impossible. At the macroeconomic level, persistent unemployment, even of a cyclical nature, is not easily reverted (Ball, 1999). Negative demand shocks raise the NAIRU, and the unemployment rate does not revert to a long-term level. Although empirical evidence on unemployment hysteresis remains weak, the extent of the 2007–09 crisis has raised some concern that part of the unemployed would de facto be excluded from the labor market, which would permanently reduce the level of potential output (OECD, 2010).
- *Creative destruction.* The impact of recessions on the demographics of firms and their innovation behavior is disputed. The Schumpeterian tradition sees recessions as productive because they hasten the attrition of the least efficient firms and contribute to creative destruction. Labor and capital freed up by bankruptcies are directed to more productive firms, which raises overall productivity. Governments should therefore not oppose the "cleansing" effect of recessions (Caballero and Hammour, 1994) by attempting to stabilize the economy. This is one of the reasons why the OECD (2010) did not expect the 2007–09 crisis to have a negative impact on TFP growth over the medium term. In contrast, another line of thinking stresses irreversible losses caused by recessions: Companies that go bust are not necessarily the least-effective ones and can simply be the most fragile or those which took more risks (Aghion et al., 2007). Furthermore, their disappearance induces a social loss because of the depreciation of capital goods and of firm-specific skills and knowledge.

Each of these arguments is theoretically relevant and the jury is out on whether output volatility is good or bad for long-term growth. Ramey and Ramey (1995) look at a sample of 92 countries and find a negative effect of GDP volatility on long-term GDP growth. For example, the "stop-and-go" policies carried out in the UK in the 1980s and 1990s are generally thought to have slowed down UK productivity growth (Barrell and Weale, 2003). Beyond this simple evidence there are, however, good reasons to believe that the sign of the relationship depends on the structure of markets. Recessions are more costly in a country where the labor market is rigid and the probability of exiting unemployment is low, or where accessing credit is difficult (so that

45. The expression was popularized in macroeconomics by Blanchard and Summers (1986). It is borrowed from physics. One speaks of hysteresis when the transformation of a material under the effect of temperature and/or pressure is irreversible: The material bears the memory of its last transformations.

firms cannot borrow to avoid going bust). Aghion and Banerjee (2005) have shown that the impact of GDP volatility on growth is more negative when financial markets are less developed.

Hence, the dichotomy between stabilization and allocation holds only as a first approximation. Macroeconomic policy is likely to have long-term effects when fluctuations are wide and market institutions do not allow economic agents to weather recessions.[46] Symmetrically, it is increasingly recognized that policies favoring long-term growth are likely to increase resilience to cyclical fluctuations.[47]

When designing a growth program, however, the interaction between short- and long-term policies is generally ignored. Here, we follow the production function sketched in figure 6.9. We start from the *A* component of the production function and discuss how policies can foster institutional improvements. We go on discussing the role of education, innovation and infrastructure policies—also affecting the *A* component. Then we look at policies addressing labor supply—the *L* component—before turning to those, such as product and financial market policies, which aim at favoring capital accumulation—the *K* component. We add a discussion on the spatial dimension of policies and conclude with a discussion on the choice of priorities.

6.3.2 Improving institutions

Imperfect as it may seem, research on institutions has produced useful policy recommendations: First, ensure that the legal framework in which the economy operates is conducive to private initiative (create an independent judiciary to enforce private contracts, fight corruption, limit red tape, ensure transparent information, etc.); second, put in place effective market regulation (create an anti-trust authority, develop proper banking regulation, ensure consumer protection, etc.); third, achieve macroeconomic stability (through, e.g., an independent central bank and stable fiscal rules, as described in chapters 3 and 4).

Those recommendations form the backbone of the agenda set out by international institutions. For example, the "New Partnership for the Development of Africa" or *NEPAD**, set up by the African Union in 2001, acknowledges the importance of so-called "good governance" for economic development, including through peer-country reviews. Likewise, when lending to low-income countries, the IMF takes their governance into account.

46. In this respect it is ironic that the US, where the labor market is very fluid and financial markets are deep, has more active stabilization policies than Europe (see chapters 3 and 4), where labor markets are more rigid and financial markets provide less insurance against macroeconomic risk.
47. The OECD has devoted a lot of attention to this issue. See notably the analysis by Drew et al. (2004) on how labor- and product-market rigidities affect the resilience of countries to temporary economic shocks.

Beyond first principles, however, it is difficult to identify a set of precise recommendations that could be used as roadmaps by governments and international organizations. Different countries rely on different institutional set-ups—as regards, for example, the role of the state in the economy—without clear impact on economic performance. In short, no single policy recipe is right for all countries and at all times.

According to Dani Rodrik (2006), the growth policy priorities at the end of the 1980s (the so-called "*Washington consensus*"*) were the following: (i) Fiscal discipline; (ii) reorientation of public expenditures; (iii) tax reform; (iv) financial liberalization; (v) unified and competitive exchange rates; (vi) trade liberalization; (vii) openness to foreign direct investment; (viii) privatization; (ix) deregulation; and (x) secure property rights. The "augmented consensus" of the early 2000s included 10 additional priorities: (xi) Corporate governance; (xii) anti-corruption; (xiii) flexible labor markets; (xiv) WTO agreements; (xv) financial codes and standards; (xvi) "prudent" capital-account opening; (xvii) nonintermediate exchange-rate regimes; (xviii) independent central banks/inflation targeting; (xix) social safety nets; and (xx) targeted poverty reduction. Since then, the financial crisis has dramatically illustrated the importance of sound regulation.

6.3.3 Investing in education, innovation, and infrastructures

Governments everywhere have an essential role in human capital accumulation, research, and infrastructure building because all three involve significant externalities. The modalities of government intervention differ across countries—some intervene directly in their financing, some indirectly through giving incentives to private agents to make growth-enhancing investments.

a) Education[48]

The rate of return from education is hard to measure, since education does not play a direct role in production. It is merely a way to transmit human capital and to reveal talent, a gift that is distributed unequally among individuals and depends on individual dispositions as well as on social assets. More precisely, it is difficult to know which part of the supplementary wage income generated by an additional year of higher education measures the marginal yield of study, pre-existing talent, or rent accruing to belonging to a particular social, ethnic or gender group.

48. We do not intend here to discuss the economics of education, as this is beyond the scope of this book. One can refer to the works of Gary Becker, Jacob Mincer, James Heckman, and others. We focus here on the link between education and the level of GDP per person.

At a macroeconomic level, however, the link between the education level and GDP per person has been well documented since the seminal study of Nelson and Phelps (1966). After controlling for other factors, Barro (2001) finds that an additional year of schooling raises medium-term growth by 0.44 percentage points. Other studies, in particular those undertaken under the aegis of the World Bank, have confirmed that (i) primary education exhibits the highest social profitability in developing countries while tertiary education is more relevant in OECD countries, (ii) the private return from education is higher than the social return because of the opportunity cost of public subsidies, and (iii) the return from female education is higher than from male education (see, e.g., Sianesi and Van Reenen, 2002, for a survey).

Education is an ideal playground for the "distance-to-the frontier" approach to economic growth outlined in the previous section. When an economy is distant from the technological frontier, in primary and secondary education it is enough to import and copy innovations found elsewhere, but those countries which approach this frontier have to invest in tertiary education to set up their own innovation capacity. Aghion et al. (2008a) find that tertiary education has a strong effect on growth in countries close to the technological frontier (whereas it does not affect growth in countries that are away from it): A one percentage point increase in the proportion of graduates in the labor force increases medium-term TFP growth by about 0.1 percentage points. This suggests that higher education is a very profitable investment in developed countries.[49]

Against this background, countries exhibit surprising disparity in their investment in human capital accumulation. Some developing countries are known for putting considerable effort into primary and secondary education (as reflected in the Human Development Index presented in chapter 1), others remain characterized by a high incidence of illiteracy. A telling example here is the comparison between Tunisia, where the female literacy rate is 65%, and Morocco where it is only 40%. Disparity can be found also among developed countries, this time in tertiary education attainment and resources invested in higher education. According to 2009 OECD statistics, less than 20% of the population aged 25–34 had reached tertiary education in 2007 in Italy—against 39% in Spain, a country of similar development level. In 2006, total expenditure on tertiary education was 1.1% of GDP (of which 0.2% comes from private funds) in Germany—against 2.9% of GDP (of which 1.9% comes from private funds) in the US, and 1.5% of GDP (of which 1.0% comes from private funds) in Japan. The very large discrepancy between Europe and the US is one of the key factors behind the lower European TFP performance documented in table 6.1.

To improve the performance of European higher education, however, money will not be enough. Research indicates that both the size of the

49. Aghion et al. (2005a) find a similar result for US states.

budget and the quality of governance enter into play in determining the research output of universities (Aghion et al., 2008). Stronger incentives for quality teaching and research at the grassroots level and increased competition between universities are needed in Europe. They do not need to imply convergence toward a single template for the financing or the governance of universities.

b) Research and innovation

Two groups of indicators are frequently used to measure research and innovation. The first group covers the effort of each country in terms of R&D spending or personnel. According to this first group of indicators, Europe as a whole lags behind Japan and the US, although there are very large discrepancies among European countries. Specifically, total (public and private) expenditure on R&D amounts to around 3% of GDP in Japan, slightly above 2.5% of GDP in the US, and slightly below 2% of GDP in the EU, with a large variance within the region (from 0.46% in Romania to 3.82% in Sweden in 2006, according to Eurostat). The Lisbon agenda includes an objective of 3% of GDP in 2010, which would bring Europe close to Japan and above the US, but it has not been effective so far.

A significant difference between Europe on the one hand and the US and Japan on the other is the contribution of privately funded R&D: It amounts to 1% of GDP in the EU against about 2% in both the US and Japan. The difference between the two sides of the Atlantic, therefore, does not come from government-funded but rather from private-sector research. The reason why European companies invest less than their European counterparts has to do with the industrial structure (the US is more specialized in research- and technology-intensive sectors), but also with market imperfections such as the relative underdevelopment of risk capital. As a consequence, cash-poor companies that would have the potential to invest in research may be constrained by the unavailability of funds (Philippon and Véron, 2008).

In addition to market imperfections, there is a broader reason for government intervention in the field of research, which is that the social return on research spending generally exceeds its private return. Many countries have introduced tax incentives for spending on R&D by companies or individuals. In the US, R&D tax credits exist both at the federal and the state levels (Wilson, 2005). The same applies in Europe, although some tax schemes have been challenged by the European Commission because they are deemed to distort markets ("State aids" in EU parlance).[50]

The second group of R&D indicators relate to outcomes, namely published articles and registered patents. As shown in figure 6.10, the EU performs

50. To know more about the Commission attitude to innovation-related tax schemes, see the *Frequently Asked Questions* section on "Tax incentives to promote R&D" on the EU Web site.

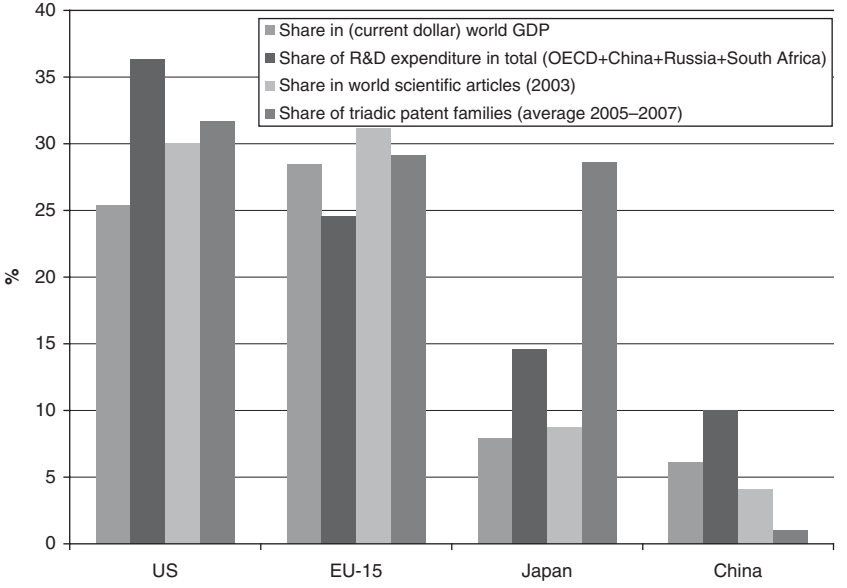


Figure 6.10 Share of world R&D effort in 2005 (in %).
Source: OECD Science, Technology and Industry Scoreboard, 2007 and 2009.

relatively well compared to its R&D efforts: Its share of world scientific articles is slightly higher than that of the US, although its share of triadic[51] patents is lower. In turn, Japan performs exceptionally well in terms of patents, although not in terms of publications. Finally, China appears as a newcomer, with a very large effort but still limited outcome.

Companies invest in research to develop new products that will give them a competitive edge or new processes that will reduce costs and improve product quality. However, every innovation is soon copied by competitors. This highlights the importance of intellectual property protection in the incentive to innovate. If a new product or a new process remains forever the exclusive property of its inventor, companies have a strong incentive to invest massively in research and development. However, the reward of innovation is appropriated by companies and their shareholders, not by consumers or the society at large. Productivity gains in other companies or sectors may be slowed down by the limitations to the dissemination of the invention. Conversely, if companies cannot appropriate the revenue of innovation because it can be accessed freely by competitors, they have little incentive to innovate. Innovation becomes a public good and it is up to taxpayers to finance it. The case of software patents illustrates this dilemma (box 6.16).

51. Triadic patents are those filed simultaneously with the US, European, and Japanese patent offices.

Box 6.16 Software Patents

Without patents, there is little incentive to innovate since any new software can be easily copied. But generalized patenting would also discourage innovation, since developers would have to pay fees on every bit and part of their new software, and for any algorithm needed to compile the code. Also, it is difficult to prove the "newness" of a software and to distinguish between technical progress and new business methods (such as the "single click" purchase patented by Amazon in the US). Finally, smaller software producers fear that large companies would tend to license any line of code as a defense against competition.

This economic dilemma is reflected by international law. The World Trade Organization agreement on *Trade-Related aspects of Intellectual Property rights* (*TRIPS*)* states that "patents shall be available for any inventions, whether products or processes, in all fields of technology, provided that they are new, involve an inventive step and are capable of industrial application." Whether software is "technology" and an "invention" is open to discussion.

In 2002–05, a highly contentious discussion developed in the EU along these lines. Software is protected by copyright but not patented as such under European law, contrary to the US and Japan. The European Patent Office (EPO) case law nevertheless views as patentable software that solves "technical problems" (as opposed to introducing new business methods). In 2002, the European Commission sought to incorporate this practice into EU law. Unsurprisingly, the proposal was supported by large firms such as Microsoft or IBM and opposed by free software and open source programmers. It was accepted by the Council of Ministers, but rejected by 648 votes to 14 in the European Parliament and therefore abandoned.

What the best regime is can only be assessed on a domain-by-domain basis. Some inventions are essentially nonrival, such as mathematical formulas and, more generally, ideas.[52] It would be absurd to hinder their dissemination. Others are essentially rival, such as manufacturing processes. Some can be replicated at low cost, such as software (box 6.16), while others cannot, such as nuclear technologies. The social value of innovation also has to be considered. Drugs are a case in point (box 6.17).

52. On the economics of knowledge and the "nonrivalry" of ideas, see Jones (2005).

Box 6.17 Fighting HIV/AIDS in Poor Countries: Public Health and Intellectual Property

Public health is a major concern in poor countries which suffer from a high prevalence of pandemics such as HIV/AIDS, tuberculosis, and malaria. Medicines to fight these diseases have been developed at a high cost by pharmaceutical companies and are, rightly, protected by patents. Such patents grant the company exclusive rights to produce and sell medicines for a long period of time, generally 20 years. As a result, the cost of therapy makes it inaccessible to many. According to the joint United Nations program on HIV/AIDS, HIV programs in low-and middle-income countries have cost US$ 13.7 billion in 2008.

Low-income countries have therefore sought to grant so-called *compulsory licenses** so that generic antiretroviral therapy could be produced locally without the consent of the patent-holder (which is, nevertheless. entitled to an adequate compensation). The WTO TRIPS agreement originally restricted generic copies to being produced mainly for the domestic market.[a] It was amended in 2003 to allow exportation of a limited list of medicines to countries that cannot produce them themselves. Under this provision, as an example, Indian pharmaceutical firms have exported generic antiretroviral drugs to African countries. In some instances, the mere threat of granting a compulsory license has led pharmaceutical companies to offer significant discounts in the local market.

The new TRIPS agreement strikes a balance between providing incentives for future inventions and disseminating more broadly existing inventions. It has been instrumental in fighting HIV/AIDS in Africa and has thus contributed to lowering mortality rates and supporting long-term GDP growth on the continent. There is, however, wide acknowledgement that intellectual property rights should remain adequately protected to allow private investment in medical research.

[a]To know more about TRIPS, see the WTO Web site, www.wto.org.

The upshot is that TFP-enhancing innovation depends on a fine balance between (i) government support and private initiative, and (ii) patent protection and the dissemination of inventions. Creating a climate that is conducive to innovation and thereby growth is the result of an elaborate chemistry.

c) Public infrastructures

Why produce goods if there is no way to bring them to the market? Economic development requires proper infrastructures such as schools, hospitals, roads,

railways, airports, dams, electricity grids, telecommunication networks, and water supply and sanitation. Such infrastructures are often financed by governments—or by foreign aid when countries are less developed—and by private money as countries grow richer and develop sophisticated financial markets. As an illustration, public investment in the four largest European countries (Germany, France, the UK, and Italy) halved from 4% of GDP in the early 1970s to 2% of GDP in the early 2000s, while it has trended upward in Greece, Ireland, Spain, and Portugal (Perée and Välilä, 2007).

In all cases, however, there is a need for government intervention:

- First, many infrastructures are natural monopolies (see chapter 1). If in private hands, the government (directly or through a dedicated agency) has to check that owners do not appropriate an excessive share of the rent they generate, and may sometimes decide that they should be accessed for free. The design of appropriate regulatory frameworks that at the same time favor competition and foster investment in infrastructure is a delicate task, especially in network industries such as telecoms, electricity, and railroads.
- Second, infrastructures involve externalities: They are used by the public at large, but they can also damage the environment. There is, therefore, a need for adequate compensation (to subsidize the gap between the private cost and the social benefit) or taxation (to compensate for damages). Dams are often controversial because they offer country-wide social benefits but cause local damage.
- Finally, there are instances in which the market cannot finance infrastructure by itself, in particular because of the lack of financial instruments to manage the risks or time horizon they are associated with. Raising money for long-term investment requires the existence of a market for very-long-term loans and bonds and for the hedging of inflation risk.

Such market imperfections may be a reason for the government to step in, but they should not be an excuse to undertake projects that have political appeal but a negative net social value. In the case of the 53-kilometer-long tunnel planned under the Alps between Lyons (France) and Torino (Italy), the social return has been estimated to be 3%, lower than the financing cost—even through riskless government debt—at the corresponding horizon.

An example of a government-sponsored infrastructure scheme is the *Trans-European networks* program launched by the European Union in 1994 in the fields of transport, energy, and telecommunications. It is funded by European governments, the EU and the European Investment Bank (the regional development bank) and involves *public–private partnerships**.[53]

53. Public–private partnerships are projects that are funded and operated through a partnership between the government and one or several private companies.

Increasingly, priority is given to communication infrastructures such as broadband internet or satellite networks. These are deemed to generate a higher social return by benefiting sectors with higher productivity gains, but there is a lack of compelling evidence. Infrastructure investments are often decided on for political reasons or on pure Keynesian grounds as a way to stabilize domestic demand. An extreme example, which involved a combination of both motives, is Japan in the 1990s (see chapter 3).

6.3.4 Increasing labor supply

The *participation rate** (the ratio between the population in the labor force and the population of active age) varies from less than 50% in Turkey to more than 80% in Iceland. In other words, if participation in the labor force (and employment) were at the Icelandic level, Turkey's income per head would be almost twice higher. This is quite an extreme example, but the variance of participation rates is nevertheless striking. As indicated by figure 6.11, there is almost no variation across countries for men between 25 and 54 years old, but there is considerable variance for women, for young workers, and for older workers.

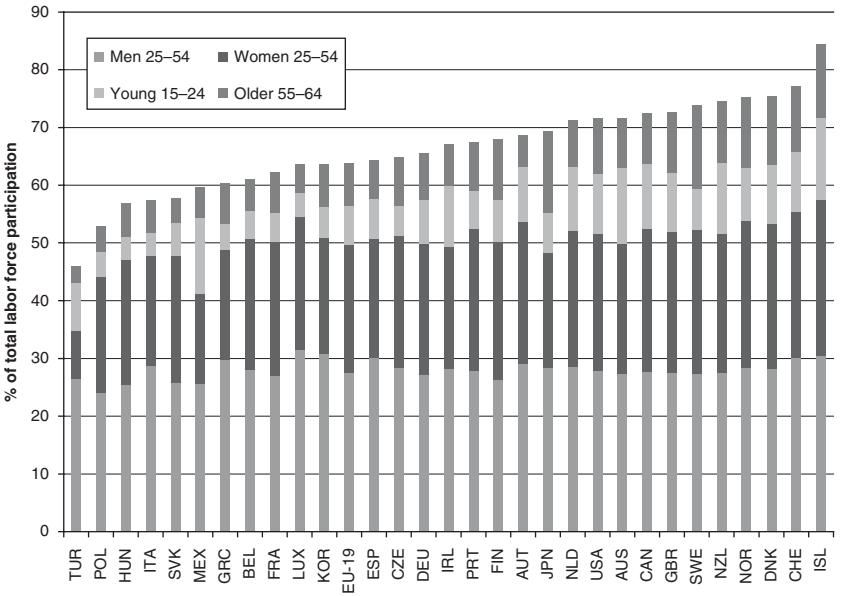


Figure 6.11 Contributions to labor-force participation rates in OECD countries, 2005.
Source: OECD Labor Force Statistics.

Reading: The figure gives the contributions to the aggregate labor-force participation rates of the participation rates of four categories: men aged 25–54; women aged 25–54; men and women aged 15–24; and men and women aged 55–64.

As for working hours (box 6.1), genuine preferences may account for some part of observed differences. But a large part of them can be ascribed to the involuntary effects of public policies. Women may be discouraged from working by taxation or because of the lack of child care infrastructure. Students may find it difficult to combine study and work because regulations do not favor part-time work. Older workers may give up working because early retirement offers a better alternative. So even though some differences may be genuine, in most countries there is room for improving public policies.

Increasing labor-force participation is especially important in industrialized countries, since the population of working age is frequently stagnant or declining while the number of pensioners is rising. In the medium term, at least, a higher participation rate can help offset the effect of aging and contribute to growth. Two main instruments have recently been used to this end:

- *In-work benefits* for unskilled workers whose labor income is only marginally above, or even below, what they can get from social programs. This is rare for full-time workers (but can happen, depending on the structure of the family), but frequent for part-time workers. To counter the disincentive effects of social benefits, several countries have introduced in-work benefit programs that help offset the effect of work that leads to losing access to means-tested benefits. Such programs include the US *Earned Income Tax Credit*, the UK *Working Families Tax Credit* and the French *Prime pour l'emploi* and *Revenu social d'activité*. The design of in-work benefits systems raises a host of difficulties, as their phasing out is itself bound to have disincentive effects.
- *Pension reforms* often include measures to improve the incentives to remain in employment, while traditional systems often involved strong incentives to leave when reaching the legal retirement age, if not before. According to the OECD, the implicit taxation on continued work in 2005 was about 50% in Belgium, France, and Japan.

Policies of this sort can help maintain a positive rate of growth of the labor force for several years. In the short term, and beyond the medium-term horizon, however, they are bound to have limited impact. The only policies that can contribute to sustaining the growth rate of the labor force in the long run are measures aimed at increasing the fertility rate and/or immigration. The fertility rate is often considered an extra-economic variable, but it can be raised by providing childcare facilities for young working families, so that labor-market participation is not an obstacle to raising children. France and Northern European countries provide examples of such policies. As for the importance of immigration, it has been understood by countries like the US, Ireland, Sweden, and the UK, where inward migration has contributed to very significant increases of the labor force and correspondingly to higher growth.

6.3.5 Making labor and product markets work better

The allocation role of markets is by no means a recent discovery, but it is fair to say that the importance of properly functioning markets has gained increasing recognition in the last two or three decades, in relation to the growing needs for factor reallocation across sectors and across firms of the same sector (see Comin and Philippon, 2005).

a) Labor markets

In an economy where labor is permanently reallocated across firms and sectors, the quality of the matching between workers and jobs becomes an important determinant of productivity and growth. First, the shorter the period during which labor remains idle after a lay-off or quitting, the higher the aggregate labor input and production. Second, the better the match between labor supply and labor demand, the higher the productivity level. Conversely, an economy where university graduates end up serving pizzas is unable to attain the productivity level that would be expected from the level of human capital.

The two objectives can be contradictory: A quick match is not necessarily a good match. This is why the performance of labor market institutions matters. In the US, there is little government involvement in the labor market and the short duration of unemployment insurance acts as a strong incentive for the unemployed to take up a new job. There is a risk that this could lead to deterioration in the quality of the matching. The magnitude of reallocations (as measured by gross flows) ensures that many opportunities exist at each point in time.

In Europe, the traditional pattern is one of job security (for those on regular contracts), but it has been undermined by changes in the structure and the dynamics of firms. It is in the Scandinavian countries that labor market institutions have undergone the deepest reforms; this has led to the emergence of a new model generally called *flexsecurity**. Workers are not offered job security anymore, but, if unemployed, they benefit from generous unemployment benefits and personalized training and placement services. Benefits are conditional on active search behavior, but they can be extended for a long period if necessary. The model is costly (expenditures on labor market policies amount to more than 4% of GDP in Denmark and 2.5% in Sweden, against 0.5% in the US) but effective in fostering quality matching. It has been adopted as a reference by the EU. Yet in practice, on-the-job protection of employees on regular contracts remains widespread in continental Europe.

b) Product markets

The functioning of markets for products and services has become increasingly prominent in the evaluation of economic performance, especially

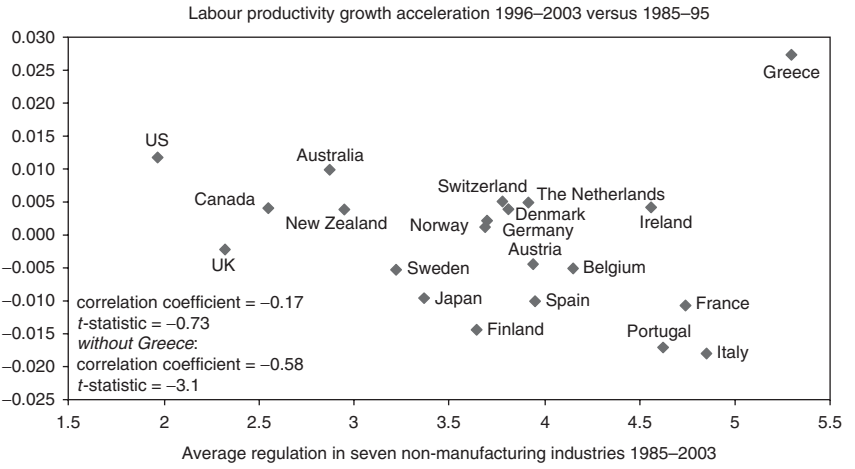


Figure 6.12 Product market regulation and labor productivity acceleration in OECD countries.
Source: Conway et al. (2006).

Reading: From 0 (least restrictive) to 6 (most restrictive).

by international institutions such as the OECD.[54] Empirical studies have demonstrated the positive impact on productivity growth of suppressing rents created by heavy-handed regulation and/or barriers to entry. Nicoletti and Scarpetta (2005) of the OECD have built synthetic indicators measuring the intensity of regulation. The value of these indicators differs widely from one country to another and this variance helps us understand the different dynamics of labor productivity (figure 6.12).

Liberalization, however, involves trade-offs. In particular, incentives to invest in research depend on the degree and nature of competition on product markets, and its role as a driver of, or an obstacle to, innovation is fiercely debated. Economists view competition as the engine of efficient resource allocation, while industrialists often accuse it of weakening industrial champions.

In early models of Schumpeterian innovation such as the canonic model of Aghion and Howitt (1992, see section 6.2), too much competition in product markets discourages innovation, since it reduces the monopoly rent that rewards it (for the same reasons, in that model, patent protection is unambiguously good for innovation). However, the case can also be made that there should be *enough* competition so that incumbent firms are challenged

54. See, for example, the annual study of the OECD on the euro area. As part of the Lisbon process, European countries also produce "structural" indicators measuring the degree of integration of markets for goods and services, openness to competition, creation and destruction of companies, etc. These indicators are available on Eurostat's Web site.

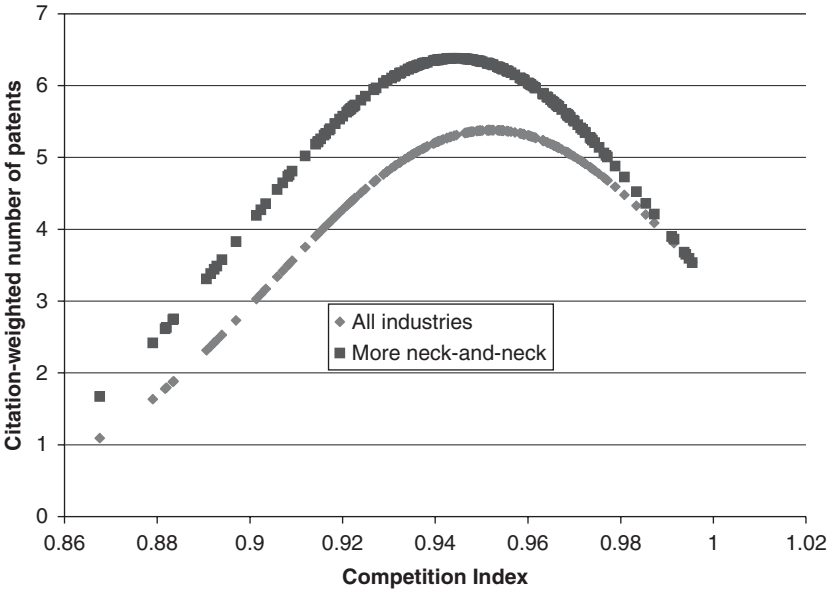


Figure 6.13 Competition and innovation in a panel of UK companies.
Source: Drawn from data communicated by Philippe Aghion. See Aghion et al. (2005b).

Reading: The figure plots a measure of competition on the *x*-axis against the citation-weighted number of patents on the *y*-axis. Competition is related to operational margins, themselves measured by operating profits net of depreciation, provisions and cost of capital, divided by sales. The relationship is supposed to be exponential quadratic and the parameters are estimated from industry-level data. "Neck-and-neck" sectors are defined as those where no company lies clearly ahead of its competitors. These sectors are more innovative for any level of product market competition.

by new entrants. Put together, this suggests an inverted-U relationship between competition and innovation. Aghion et al. (2005b) have provided a theoretical underpinning for this trade-off and have uncovered this inverted-U pattern in industry-level data by relating the number of patents submitted by UK companies to a measure of competition based on operational margins (figure 6.13).

The controversy raised by the Microsoft case is an illustration of the "inverted-U" pattern. Competitors filed cases against Microsoft, complaining that the company was attempting to obstruct them by abusing its dominant position on the market (e.g., by bundling its Windows operating system with other software such as Windows Media Player), and that this would stifle innovation (the left-hand tail of the inverted-U curve). Microsoft argued that constraining its ability to make profits would damage its innovation

capacity (the right-hand tail of the curve) and eventually hurt consumers. It was fined by the European Commission and lost its appeal in 2007. The court considered that more competition would benefit innovation—pointing to the left-hand-side of the inverted-U curve.

The debate is especially acute as regards network industries. In the telecoms sector, for example, liberalization has led to the entry of new players who have challenged the former monopolies. The latter were helped by asymmetric regulation aiming at preventing predatory pricing and other forms of lethal retaliation by the incumbents. However, as competition developed, there was a need to ensure that incentives to innovate remained strong enough. In the words of a former head of the French telecom regulation agency:

> On the one hand, a lightly regulated monopoly is likely to be too prudent and slow, and to under-invest. On the other hand, heavy-handed regulation might lead to insufficient profitability for new investment and risk-taking. There is a way out of such a dilemma: regulation has to evolve along the successive investment cycles. A strong regulation on infrastructures inherited from the monopoly situation gives the new entrants the opportunity to climb the investment ladder and curtails the competitive advantage which the former monopoly derives from the ownership of these infrastructures. Then, there is an incentive for the former monopoly, the incumbent, to invest in order to restore part of this advantage. This incentive is enhanced if all the actors can participate in the new investment cycle in a context where the regulation becomes lighter.
>
> Paul Champsaur (2008)

More generally, Griffith and Harrison (2004) have shown that reforms which have facilitated market entry and reduced administrative costs in Europe have led to lower profit margins and have supported investment and employment.[55] However, governments have to make sure that markets deliver appropriate price and quality signals and that competition is not stifled by collusion among existing players. This is what market regulation is about. Depending on the legal system, it is enforced by independent anti-trust agencies and/or by courts, and by specialized, technical agencies such as the US Federal Communication Commission and Food and Drug Administration.

6.3.6 Developing and regulating financial markets

Many growth strategies tend to overlook the role of financial markets. This is the case in Europe, where financial integration is often regarded as a goal in itself and where the Lisbon agenda is almost silent on the contribution of financial markets. This neglect is not justified (see also chapter 8).

55. Griffith and Harrison also find that insufficient profits would be unfavorable to R&D, consistent with the existence of a downward-sloping branch on figure 6.9, but they take this result with a grain of salt. See Schiantarelli (2005) for a survey.

From a macroeconomic standpoint, the role of the financial system is threefold: It transfers income over time, favoring intertemporal behavior; it collects household savings and directs it to finance the accumulation of capital, at home or abroad; and it helps individuals and companies shed risks they do not want to bear. It therefore fulfils a key allocation function in the sense given in chapter 1.[56] Three channels of influence on long-term growth can be identified (Pagano, 1993):

- *Lower cost of capital.* Collecting household savings entails transaction costs, which reflect the costs of production of financial services, but also the taxes and regulations in force and oligopoly rents. Competition in the financial sector increases the effectiveness of the intermediation process and lowers the cost of capital.
- *Higher savings.* By giving confidence to savers, a robust financial system makes it possible to increase the saving rate, thus GDP per person (this is again the Solow–Swan model of section 6.2).
- *Better allocation of capital.* The financial system makes it possible to collect and share information on investment projects, to diversify risk and to finance innovation: In a nutshell, to direct saving toward the most productive projects.

Let us take each of these three channels in turn.

a) The cost of capital

The Ramsey model of section 6.2 tells us that there are instances in which an economy has too much capital, yielding an insufficient return to support consumption. Table 6.1 in section 6.1 suggests that the situation of Europe is rather the opposite. The lower GDP per person seems at least partly due to an insufficient capital/labor ratio.

Public policies can theoretically affect the cost of capital through monetary, fiscal, regulatory, and tax policies. Monetary policy once attempted to lower interest rates to encourage investment, but in the present context of developed financial markets, its influence on long-term interest rates is very limited (see chapter 4). Similarly, fiscal policy hardly affects the interest rate, since capital flows freely across countries. As regards regulatory policies, they have lost the impact they had when governments could direct household savings to the financing of corporate investment or to specific sectors. The main available instrument is therefore tax policy, especially the taxation of corporate earnings.

It has not always been so. Eastern Asian economies maintained until the early 1990s a system of *financial repression**, i.e., interest rates maintained

56. See, for example, Greenwood and Jovanovic (1990), Levine (2005) and the pioneering works of Schumpeter (1911) and of Gurley and Shaw (1955).

at low levels by governments to encourage investment.[57] Similarly, German firms long benefited from an artificially low cost of capital thanks to relationship banking and public guarantees granted to regional banks. And many governments still extend subsidized loans to help some sectors, such as agriculture. The EU generally views such interventions, deemed *state aids**, as obstacles to free competition which should be prohibited unless for the sake of general interest.

Turning to taxes, temporary tax exemptions have a limited effect on capital expenditures, save for their timing, but the permanent features of corporate taxation such as amortization schemes, the definition of the tax base, and headline tax rates do play a role as incentives or disincentives to invest (see chapter 7).

b) The level of savings

If capital moves freely across countries, then capital expenditures are not constrained by the availability of domestic savings, and there is no point in inciting households to save more. But for various reasons discussed in chapter 5, savings and investment remain (weakly) correlated and governments continue to have recourse to policies aimed at encouraging savings. An example is pension reform. Funded pensions are a form of forced savings and help increase GDP per person, provided that pension money is invested in corporate bonds or stocks (but they can encourage excessive investment if savings are already high).

Policies aimed at encouraging savings suffer from several drawbacks. An obvious one is the short time horizon of policymakers, who are bound by the political cycle. Impatient policymakers stress consumption at the expense of savings. Ignorant ones stress both. Another drawback is the difficulty of judging the adequate level of capital: At an aggregate level, this is about dynamic inefficiency (i.e. over- or under-capitalization) in the sense of the Ramsey model; at the industry level, this is about whether companies are constrained by insufficient capital supply or rather by a poor demand outlook or bad functioning of the labor market.

The relation between financial development and saving rates is more ambiguous than it may seem. In developed countries, household access to insurance through access to financial markets, and portfolio diversification reduces their precautionary saving. It also lifts their financing constraint by giving them access to financing instruments such as residential mortgages. The low level of household savings in the US and in the UK is a

57. This policy indeed contributed to promoting domestic investment, but may also have led to overinvestment rather than total factor productivity (see Young, 1992). In their seminal works, McKinnon (1973) and Shaw (1973) argue that financial repression is altogether a barrier to successful economic development. Both the theoretical and the empirical literatures, notably in the wake of severe financial crises, have subsequently qualified the McKinnon and Shaw hypothesis.

case in point. However, in the first stage of the development process, a well-functioning financial system contributes to directing savings toward productive investment.

c) The allocation of savings

Finally, governments can also deal with the allocation of savings. *Credit rationing**, which was in place until the 1980s in many developed economies, amounted to a central planning of capital allocation to individual companies. Fiscal incentives can nowadays still channel savings—for example, to finance innovation through R&D tax credits (see above) or to favor small- and medium-size enterprises. Tax policy and regulatory policies influence decisions to invest in equity or bonds. In 2000, the proportion of shares and mutual funds (predominantly invested in equity) in the financial wealth of households was 31% in Europe, 46% in the US, and 15% in Japan according to Babeau and Sbano (2002).

Until the financial crisis of the late-2000s, the ability of the US financial market to innovate and channel funds to the most productive uses was regarded as a major competitive advantage of the US economy in comparison to Europe and Japan. Former Federal Reserve Chairman Alan Greenspan has described the role of the US financial system in the so-called "new economy" of the 1990s:

> Our financial system, whose job it is to ensure the productive use of physical capital, has been such a crucial part of our overall economy, especially over the past two decades. It is the signals reflected in financial asset prices, interest rates, and risk spreads that have altered the structure of our output in recent decades toward a different view of what consumers judge as value. . . .
>
> Clearly, our high financial returns on investment are a symptom that our physical capital is being allocated to produce products and services that consumers particularly value. A machining facility that turns out an inferior product or a toll road that leads to nowhere will not find favor with the public, will earn subnormal or negative profits, and in most instances will exhibit an inability over the life of the asset to recover the cash plus cost of capital invested in it.
>
> Thus, while adequate national saving is a necessary condition for capital investment and rising productivity and standards of living, it is by no means a sufficient condition.
>
> Alan Greenspan (1998)

The financial crisis has led to reconsideration of the usefulness and dangers of the financial innovations of the 1990s and 2000s, beginning with securitization and leverage (Chapter 8), and therefore Alan Greenspan's diagnosis that: "clearly, our high financial returns on investment are a symptom that our physical capital is being allocated to produce products and services that consumers particularly value," but it does not

question the importance of a well-functioning financial system for long-term growth.

Thomas Philippon and Nicolas Véron (2008) have pointed out that a major development of the US financial system in recent decades has been the rise of corporate finance, and they have highlighted the rising share in total corporate investment of low-cash firms that need to rely heavily on external finance to innovate and invest. This suggests that the financial system has effectively contributed to the emergence of a growth model that relies on the entry and the fast rise of new players that, bringing to the market new products and productivity-enhancing technologies, are able to challenge the incumbents. As a consequence, only three European firms that were established after 1975 belonged to top 500 listed companies in the world, against 26 US firms.

d) Implications

As illustrated by the financial turmoil of 2007–09 (and many previous crises), however, financial markets can also propagate risks and deter investment. To fulfill their growth-enhancing role, they need solid institutions: Independent, technically educated regulators who can follow the pace of financial innovation, prevent excessive risk-taking by financial intermediaries and make sure that their capital base is broad enough to cushion temporary losses; laws protecting savers from financial product "misselling" so as to maintain confidence in the financial system; supervision of markets to ensure their ability to produce accurate accounts and nonfragmented and transparent asset prices.

Developing the financial sector is a means, not an end. In the 1990s, the IMF and the OECD wrongly believed that the benefits of financial development would materialize automatically, and they forced financial openness upon reluctant emerging-market countries. Korea's opening of financial markets is a case in point: It was part of the "accession package" to the OECD in 1996 and in 1997 suffered a financial crisis. The 1990s and 2000s were an era of painful and unfinished apprenticeship to the requirements of financial liberalization. This includes the macroeconomic framework and exchange-rate policy, but also the sequencing of reforms and prudential regulation. In addition, the benefits of opening the financial sector to foreign competition are all the more important, since a country is not big enough to create economies of scale and sufficient risk diversification. Finally, financial development has to be gradual, in view of the risks of instability inherent in modern finance.

In Europe, the introduction of the euro has accelerated cross-border integration of financial markets. It has created a unified monetary market and closely integrated markets for government and corporate bonds and wholesale financial services. Asset management, retail banking, and SME financing have remained fragmented. This fragmentation precludes competition, prevents economies of scale from materializing in the financial industry, hurts small- and medium-sized enterprises that do not have access to global capital

markets, and hampers the diversification of risks. In short, it favors rent-seeking by financial institutions at the expense of European households and companies.

Integrating financial services ranks high on the EU agenda since the European Commission and EU Member States have drafted a "Financial Service Action Plan" (European Commission, 2005) aimed at harmonizing the regulation on financial products, consumer protection, and the functioning of the markets. The EU regulatory and supervisory framework has been further streamlined after the financial crisis (chapter 8).

Market participants can take excessive risks and transfer them to the households, directly through the origination and distribution of complex financial assets, or indirectly when financial institutions are bailed out by taxpayers. Financial markets therefore need very solid institutions. These are: Independent and technically knowledgeable controllers, regulatory and supervisory frameworks that are not outpaced by financial innovation, crisis management rules that do not create moral hazard, and proper incentive structures within financial institutions (e.g., compensation schemes, risk control, and compliance rules) and other actors such as rating agencies. The 2007–09 crisis has shown that these institutions were in many respects defective in the world's most developed economies, prompting government intervention to strengthen the regulatory framework so as to "save capitalism from capitalists," to use an expression coined by Rajan and Zingales (2003). This is further discussed in chapter 8.

6.3.7 Countering the effects of distance and history

So far we have envisaged growth policies mostly at the level of a country. However, countries or supranational entities like the EU also implement regional development policies with the aim of fostering growth.

Those are, in principle, distinct from mere redistribution policies. Regions (and cities) specialize dynamically according to their comparative advantages with capital, and to a lesser extent labor, being permanently relocated across regions. The combination of history, geography, and market forces usually results in a very uneven distribution of income and wealth.[58] Inequality can to some extent be corrected through tax-based redistribution, but the real issue is whether policy can foster growth in the less-developed regions. This is the aim of regional policies. In Europe, for example, *structural funds** top-up country-level redistribution schemes and focus on growth-enhancing investments. Structural funds, which are cashed in by regions,[59] amount to 31% of the EU

58. As an example, in 2005, individual income in European regions ranged between 2519 euros per year on average in North-eastern Romania and 76053 euros per year on average in Inner London, a one-to-thirty ratio.

59. One-tenth of it, known as the "Cohesion Fund," is distributed at a country level.

federal budget, with an aim of equalizing GDP per person across European regions.

Are such policies economically efficient? Neoclassical growth theory envisages the convergence of regional income per person, conditional to their level of human capital or to the quality of their institutions. European data do not exhibit unconditional convergence: Convergence is a fact among EU countries but not among EU regions (OECD, 2004, Part 2), and fast-converging regions have been rich regions in poorer countries, such as Catalonia.

Economic geography has shed a new light on the discussion of regional policies.[60] As seen in section 6.2 (and, in particular, in the "core/periphery" model of box 6.12), there is a tension between *agglomeration forces* (a.k.a. *polarization forces*) arising from positive spillovers between economic activities, and *dispersion forces* resulting from transport costs and possibly congestion costs (such as traffic jams or water pollution in big cities). A casual glance at figure 6.14 confirms the power of agglomeration forces. Most of the EU's richest regions lie along a central axis which goes from Northern Italy to Germany, The Netherlands, then to London. Most of the US's richest states lie along the nation's coasts and borders. Le Gallo and Dall'erba (2006) have studied the time dynamics of regional wealth in Europe and uncovered a strong dependence of the convergence process of a region's GDP per person on the wealth of its neighbors, leading to the formation of a "nonconvergence club" of peripheral regions.

Agglomeration forces have strong policy implications. In a world where they are dominant, inequality among regions increases over time, but this is the outcome of an economically efficient process which leads to a higher GDP growth *at the aggregate level* by exploiting the positive spillovers which arise when activities are clustered. Governments can operate redistribution across regions using (preferably lump-sum) tax transfers, but they should not oppose agglomeration; quite the contrary: "Catalytic" intervention to create industry clusters is welcome. Such catalytic intervention is more effective in new and R&D-intensive industries where existing capital plays a less important role. Conversely, public incentives are less effective in industries where spillovers are already very strong. Strong public support for the Frankfurt and Paris financial centers in the 2000s could not threaten the dominance of London as the European marketplace, given the strength of agglomeration forces and the spillovers arising from a better access to the global pool of highly qualified, English-speaking labor.

Why not improve transport infrastructures so as to rebalance agglomeration and dispersion forces and support peripheral regions in an economically effective way? The consequences can be unintended, since reduced transportation costs can, at least in a first stage, encourage concentration by easing the relocation of the labor force. In France, high-speed trains have encouraged

60. See Martin (1999) and Baldwin et al. (2003, ch. 17) for a detailed discussion.

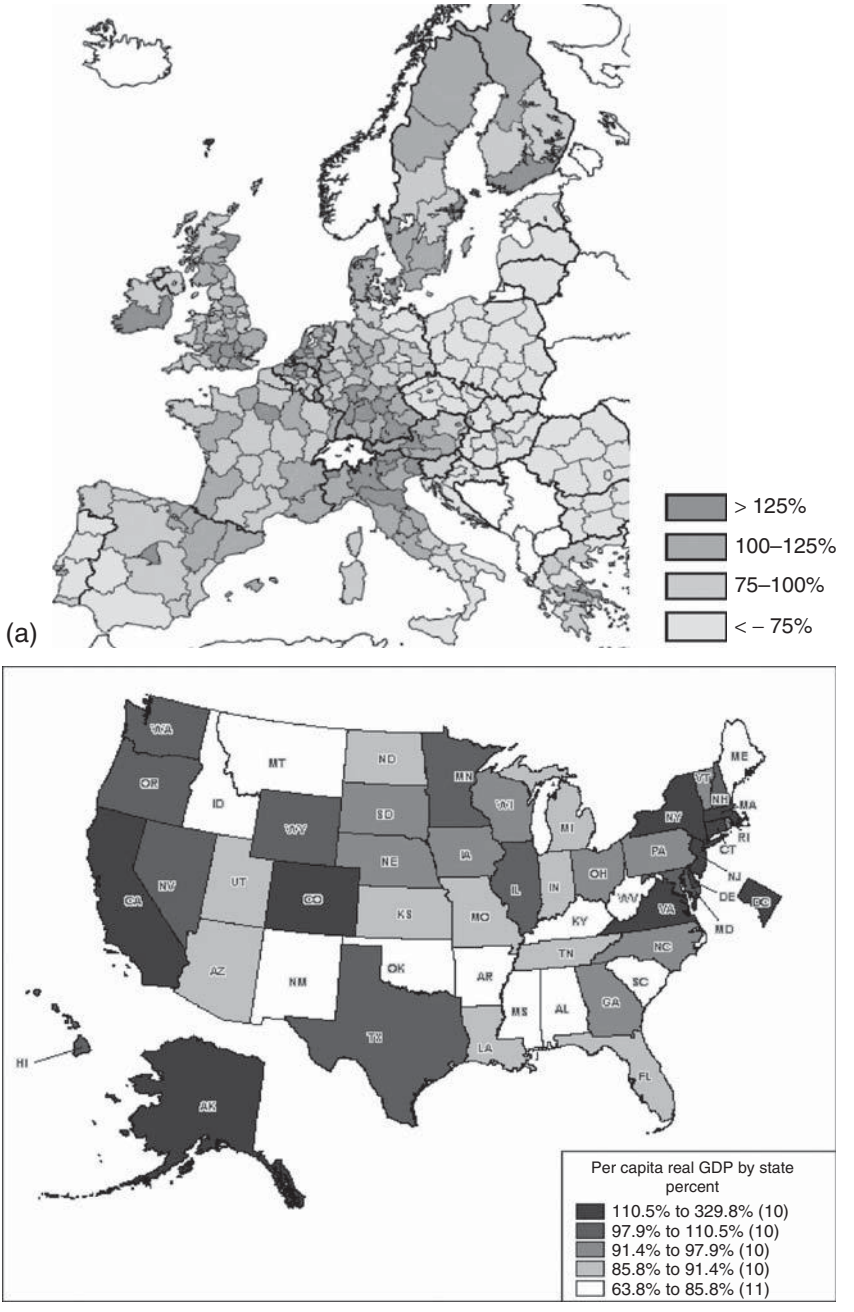


Figure 6.14 GDP per person in EU and US regions. a) EU-25, b) US.
Source: Eurostat and US Bureau of Economic Analysis.

Reading: Purchasing power standard as a % of EU-25 average in 2003, and real GDP as a % of US average in 2006.

concentration in the Paris region by making it easier to live in remote places and work in the capital. This is economically efficient (because it increases GDP in the aggregate) but geographically unequal. The remedy is to increase fiscal transfers to peripheral regions. In the EU context, this is the essence of the regional policy reform proposed by Sapir et al. (2004), who advocate better identifying the allocation and redistribution functions by setting up a "growth fund" at an EU level and a "convergence fund" devoted to less-advanced regions.

A more forward-looking way to tackle geographical inequalities is to encourage the diffusion of ideas and knowledge so that peripheral regions can "jump" to a more human-capital-intensive, less physical-capital-intensive development regime (Martin, 1999). This implies, for instance, investing in mobile-phone and broadband-internet access. The rise of the Indian ICT industry can be understood in that context: Fast development of electronic communication infrastructures has supplemented largely defective transport infrastructures. According to the World Bank, 17% of the population of sub-Saharan Africa had a mobile phone in 2006 as compared to less than 1% in the 1990s. As network coverage expands, and the price of internet access goes down, social and banking services can be provided over mobile networks and increase total factor productivity.

As seen in section 6.2, endogenous growth models also highlight the possibility of multiple equilibriums and the role of history in shaping growth trajectories. As a result, many countries or regions face the challenge of devising policies to escape low-development traps. There are two kinds of complementary remedies:

- Open up the domestic economy to international markets to reap the productivity benefits of specialization without constraining the consumers' choice. Such strategies were inaugurated by Britain's repeal of the protectionist Corn Laws and embracing free trade in the 1830s, which proved vastly beneficial to its growth. Yet the empirical literature on trade opening and economic growth does not reach firm conclusions (Rodriguez and Rodrik, 2000). In short, no country has reached a sustainable high-growth path without opening up to trade; yet trade opening does not suffice to generate growth. A reason for this result is that trade opening may also push an economy in the direction of specializing in traditional sectors in which it has comparative advantage, such as agriculture, which may hamper long-term productivity growth. While nineteenth century Britain was embracing free trade, it was also benefiting from the industrial revolution.
- Convince economic agents that future development justifies investing today. Krugman (1991a) and Murphy et al. (1989) have modeled situations where industrial development is not deterministic because it depends on demand expectations. Both take-off and stagnation are possible, depending on initial expectations. Underdevelopment as a

coordination failure was the dominant model in the postwar years[61] and echoes the "big-push" theory of development that was very popular in the 1950s and 1960s. In this context of multiple equilibriums, the capacity of governments or international institutions to influence expectations and help move the economy from a particular equilibrium to another one becomes crucial. In this coordination role, governments need to be credible (in the same way as they need credibility for the management of short-term demand as discussed in chapters 3 and 4). Their credibility can be backed by kick-starting productivity-enhancing investment with public money, by engaging in an overhaul of the regulatory and tax systems, or by seeking public support of international organizations such as the OECD and the IMF to their reform strategy.

However, forceful, "big-push" strategies crucially depend on expectations and are therefore inherently fragile. This was illustrated by the failure of the forced industrialization strategies of many developing countries in the 1960s. Once in place, newly created industries have to generate lasting TFP gains, which brings us back to the preceding set of recommendations. It was not the 1950s "Great Leap Forward" that ensured China's economic take-off, but its transition to a market economy and its opening to international trade in the 1990s.

6.3.8 Choosing priorities

After squeezing theory as hard as we could to extract its policy consequences, we still do not have a recipe for long-term growth. The renewal of growth theory in the 1990s has made it much richer, and has improved its ability to match empirical data. However, one should acknowledge that the link between policies and outcomes is more tenuous than for the policies outlined in the previous chapters, if only because the time-scale is much more extended and mechanisms are much more complex. Governments should not take these difficulties as an excuse to focus on short-term growth only. They are already all too tempted to do so, given their short political tenures. For developing countries as well as for Europe, investing in growth is crucial.

Successful growth strategies require the identification of priorities. Among the many factors with a bearing on long-term growth, governments need to choose a few on which to focus—because political capital is always scarce. How to identify those factors has been the topic of recent research by the OECD and by a group of Cambridge economists (box 6.18). There is much to criticize in these approaches but their merit is to focus on the selection of a few policy areas where action can deliver results.

61. See, for example, Ray (2001) and Krugman (1994a). Initial work on the subject can be traced to Young (1928), and in particular to Rosenstein-Rodan (1943).

In selecting priorities, governments also take into account political constraints, such as the distribution of winners and losers from the reforms and their prospective voting behavior. In effect, political economy is key to understanding why growth-enhancing reforms are often not implemented, in spite of their potential effects on welfare. How policy design can help in alleviating those constraints and contribute to building support for growth is a major issue in developed as well as developing countries. Policy design and the quality of the policy-making process, therefore, play a crucial role. As highlighted by the UN-commissioned Spence Report:

> A country's fortunes depend on stopping bad policies as well as implementing good ones. Fallacies and follies must be identified, criticized, and rejected. . . . Successful countries owe a lot to an environment in which all ideas, good and bad, are exposed to review and vigorous debate.
>
> Commission on Growth and Development (2008)

Box 6.18 Alternative Methods for Identifying Policy Priorities

The OECD Method

In its *Going for Growth* report series, the OECD proposes a method for selecting country-by-country policy priorities.

The starting point is the level of GDP per person vis-à-vis the US—chosen as the *numéraire* to assess relative performance. This GDP-per-person performance is next broken down into two performance indicators, for labor-resource utilization and for productivity. Then,

- The labor-utilization gap is further decomposed into three performance indicators: The average number of hours worked per employee, total employment as a ratio of working-age population, and the share of the working-age population in the total population.
- The labor-productivity gap is further broken down into two performance indicators: Capital services available per hour worked and total factor productivity.

To identify policy levers, the OECD starts from a series of 50 policy indicators covering labor markets, product markets, taxation, education, and research. Cross-country evidence serves to identify which policy indicators best explain divergence across countries for the performance indicators.

The identification of policy priorities for each country is finally based on matching poor policy settings and weak performance. Policy areas selected for action are those where a country's policy appears to be relatively weak and which correspond to poor performance. For example, a high rate of implicit taxation on continuing work only leads to selecting pension

reform as a key priority if the country is characterized by a low level of participation for older workers. In the end, the OECD selects three indicators-based priorities for each country.

The Growth Diagnostics Method

In a 2008 paper, Ricardo Hausmann, Dani Rodrik and Andres Velasco propose a method for deciding on priorities. They assume that obstacles to growth can be regarded as a series of distortions that introduce wedges between the private and social values of a series of activities (e.g., investment, labor supply, human capital accumulation, etc.). Removing any of those distortions through structural reforms has both direct effects and general equilibrium effects on all activities. The growth-diagnostics methodology proposes to rank reforms according to their direct effect and to start with the reforms whose direct effects are strongest. In practice, this involves following a top-down decision tree, and determining in each case what prices tell us about the effect of distortions. For example, high private returns on education suggest that lack of education is likely to be a severe constraint on growth.

Obviously, the best strategy would be to remove all distortions at once. But Hausmann et al. (2008) argue that this is not a realistic option and that, in practice, governments often start with the easiest reforms or simply "do as much as they can." Such approaches are wrong in theory and unlikely to deliver results in practice, as is the idea that one should start with the largest distortions. The authors also dismiss sophisticated second-best strategies that take into account interactions as marred with excessive uncertainty.

References

Acemoglu, D., S. Johnson, and J. Robinson (2001), "Colonial Origins and Comparative Development: An Empirical Investigation," *American Economic Review*, 91, pp. 1369–401.

Acemoglu, D., P. Aghion, and F. Zilibotti (2002), "Distance to Frontier, Selection, and Economic Growth," *Journal of the European Economic Association*, 4, pp. 37–74.

Acemoglu, D., S. Johnson, and J. Robinson (2004), "Institutions as the Fundamental Causes of Long-Run Growth," in Aghion, Ph., and S. Durlauf (eds.), *Handbook of Economic Growth*, Elsevier, chapt. 6, pp. 385–472.

Aghion, P., and A. Banerjee (2005), *Volatility and Growth*, Clarendon Lectures in Economics, Oxford University Press.

Aghion, P., and P. Howitt (1992), "A Model of Growth Through Creative Destruction," *Econometrica*, 60, pp. 323–51.

Aghion, P., L. Boustan, C. Hoxby, and J. Vandenbussche (2005a), "Exploiting States' Mistakes to Identify the Causal Impact of Higher Education on Growth," mimeo, Harvard University.

Aghion, P., B. Bloom, R. Blundell, R. Griffith, and P. Howitt (2005b), "Competition and Innovation: An Inverted-U Relationship," *The Quarterly Journal of Economics*, 120, pp. 701–28.

Aghion, P., P. Askenazy, N. Berman, G. Cette, and L. Eymard (2007), "Credit Constraints and the Cyclicality of R&D Investment: Evidence from France," mimeo, August.

Aghion, P., P. Askenazy, R. Bourlès, G. Cette, and N. Dromel (2008a), "Éducation supérieure, rigidités de marché et croissance," in Aghion, P., Cette, G., Cohen, E., and Pisani-Ferry, J. (eds.), *Les leviers de la croissance française*, Report to the French Council of Economic Analysis no. 72.

Aghion, P., M. Dewatripont, C. Hoxby, A. Mas-Colell, and A. Sapir (2008b), *An Agenda for Reforming European Universities*, Bruegel, Brussels.

Alesina, A., and D. Rodrik (1994), "Distributive Politics and Economic Growth," *The Quarterly Journal of Economics*, 109, pp. 465–90.

Arnold, J.M., A. Bassanini, and S. Scarpetta (2007), "Solow or Lucas?: Testing Growth Models Using Panel Data from OECD Countries," *OECD Working Paper* 592.

Arrow, K. (1962), "The Economic Implications of Learning By Doing," *Review of Economic Studies*, 29, pp. 155–73.

Asian Development Bank (2007), *Key Indicators 2007: Inequality in Asia*, Manila: Asian Development Bank.

Askenazy, P. (2001), *Des 35 heures à la nouvelle economie. Changements organisationnels et diffusion de l'innovation.* Notes de l'Ifri 27, Paris: La Documentation Française for the French Institute of International Relations (Ifri).

Attali, J. (2008), *Rapport de la commission pour la libération de la croissance française*, Paris: La Documentation Française.

Babeau, A., and T. Sbano (2002), "Household Wealth in the National Accounts of Europe, the United States and Japan," Statistics Working Paper 2003:2, Paris: Organisation for Economic Cooperation and Development.

Baldwin, R., R. Forslid, P. Martin, G. Ottaviano, and F. Robert-Nicoud (2003), *Economic Geography and Public Policy*, Princeton University Press.

Ball, L. (1999), "Aggregate Demand and Long-Run Unemployment," *Brookings Papers on Economic Activity*, no. 2, pp. 189–251.

Banerjee, A., and E. Duflo (2003), "Inequality and Growth: What Can the Data Say?," *Journal of Economic Growth*, 8, pp. 267–99.

Barrell, R., and M. Weale, (2003), "Designing and Choosing Macroeconomic Frameworks: The Position of the U.K. after Four Years of the Euro," *Oxford Review of Economic Policy*, no. 19, pp. 132–48.

Barro, R. (2001), "Human Capital and Growth," *American Economic Review*, 91, pp. 12–17.

Barro, R., and X. Sala-ì-Martin (1991), "Convergence Across States and Regions," *Brookings Papers on Economic Activity*, The Brookings Institution, pp. 107–82.

Barro, R., and X. Sala-ì-Martin (1995), *Economic Growth*, McGraw-Hill.

Barterlsman, E., S. Scarpetta, and F. Schivardi (2003), "Comparative Analysis of Firm Demographics and Survival: Micro-Level Evidence for the OECD Countries," *OECD Economics Department Working Papers* 348, Paris: Organisation for Economic Cooperation and Development.

Basu S., J.G. Fernald, N. Oulton, and S. Srinivasan (2004), "The Case of the Missing Productivity Growth, or, Does Information Technology Explain Why

Productivity Accelerated in the US but not the UK?," in Gertler, M., and K. Rogoff (eds.), *NBER Macroeconomics Annual 2003*, 18, The MIT Press, pp. 9–82.
Benabou, R. (1996), "Inequality and Growth," in Bernanke, B., and J. Rotemberg (eds.), *NBER Macroeconomics Annual 1996*, 11, The MIT Press, pp. 11–74.
Blanchard, O. (2004), "The Economic Future of Europe," *Journal of Economic Perspectives*, 18, pp. 3–26.
Blanchard, O., and L. Summers (1986), "Hysteresis and the European Unemployment Problem," in S. Fischer (ed.), *NBER Macroeconomic Annual 1986*, 1, Cambridge: National Bureau of Economic Research, pp. 15–90.
Borner, S., F. Bodmer, and M. Kobler (2003), *Institutional Efficiency and its Determinants: The Role of Political Factors in Economic Growth*, OECD Development Centre Studies, Paris: Organisation for Economic Cooperation and Development.
Bourguignon, F., and C. Morrisson (2002), "Inequality Among World Citizens: 1820–1992," *American Economic Review*, 92, pp. 727–44.
Braudel, F. (1981–84), *Civilization and Capitalism, 15th–18th Centuries*, 3 vols. (*The Structures of Everyday Life*, *The Wheels of Commerce*, *The Perspective of the World*), Harper & Row, original editions in French (1979).
Braudel, F. (1985), *La dynamique du capitalisme*, Arthaud.
Caballero, R., and M. Hammour (1994), "The Cleansing Effect of Recessions," *American Economic Review*, 84, pp. 1350–68
Cette, G. (2004), "Productivité et croissance: diagnostic macroéconomique et lecture historique," in Cette, G., and P. Artus, *Productivité et croissance*, rapport au Conseil d'Analyse Economique 48, Paris: La Documentation Française.
Cette, G., J. Mairesse, and Y. Kocoglu (2004), "Diffusion des TIC et croissance potentielle," *Revue d'Économie Politique*, 114, janvier–février, pp. 77–97.
Chamberlin, E. (1933), *Theory of Monopolistic Competition*, Harvard University Press.
Champsaur, P. (2007), "Competition, Innovation and Investment in Network Industries," Presentation at the IDEI/Bruegel conference on Competition, Innovation and Investment in Network Industries, Brussels, 26 October.
Collier, P. (2007), *The Bottom Billion: Why the Poorest Countries Are Failing and What Can Be Done About It*, Oxford University Press.
Combes, P.-P., T. Mayer, and J.-F. Thisse (2006), *Economie géographique—L'intégration des regions et des nations*, Economica.
Comin, D., and T. Philippon (2005), "The Rise in Firm-level Volatility: Causes and Consequences," *NBER Macroeconomics Annual*, 20, The MIT Press, pp. 167–228.
Commission on Growth and Development (2008), *The Growth Report. Strategies for Sustained Growth and Development*, Washington DC: The World Bank.
Conway, P., D. de Rosa, G. Nicoletti, and F. Steiner (2006), "Regulation, Competition and Productivity Convergence," *OECD Economics Department Working Papers*, no. 509, Paris: Organisation for Economic Cooperation and Development.
Coulombe, S. (2007), "Globalization and Regional Disparity: A Canadian Case Study," *Regional Studies*, 41, pp. 1–17.
Coulombe, S., and F.C. Lee (1995), "Convergence across Canadian provinces, 1961 to 1991," *Canadian Journal of Economics*, 28, pp. 155–78.
David, P. (1985), "Clio and the Economics of QWERTY: The Constraints of History," *American Economic Association Papers and Proceedings*, 75, pp. 333–37.
David, P. (1990), "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox," *American Economic Review*, 80, pp. 355–61.

Deininger, K., and L. Squire (1996), "A New Data Set Measuring Income Inequality," *World Bank Economic Review*, 10, pp. 565–91.
Diamond, J. (1997), *Guns, Germs and Steel. The Fate of Human Societies*, W.W.Norton and Company.
Dixit, A., and J. Stiglitz (1977), "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, 67, pp. 297–308.
Domar, H. (1946), "Capital Expansion, Rate of Growth and Employment," *Econometrica*, 14, pp. 137–47.
Drew, A., M. Kennedy, and T. Sløk (2004), "Differences in Resilience between the Euro-area and US Economies," *OECD Working Paper*, no. 382, Paris: OECD.
Easterly, W. (2001), *The Elusive Quest for Growth*, The MIT Press.
European Commission (2005), *Financial Services Policy, 2005–2010*, White Paper, Brussels: European Commission.
Fogel, R. (2007), "Capitalism and Democracy in 2040: Forecasts and Speculation," *NBER Working Paper* 13184, Cambridge, MA: National Bureau of Economic Research.
Galor, O., and D. Tsiddon (1997), "Technological Progress, Mobility and Economic Growth," *American Economic Review*, 87, pp. 363–82.
Gancia, G., and F. Zilibotti (2004), "Horizontal Innovation in the Theory of Growth and Development," in Aghion, P., and S. Durlauf (eds.), *Handbook of Economic Growth*, Elsevier, Chapt. 3, pp. 111–70.
Gordon, R. (2000), "Does the 'New Economy' Measure up to the Great Inventions of the Past," *Journal of Economic Perspectives*, 14, pp. 49–74.
Gordon, R. (2003), "Exploding Productivity Growth: Context, Clauses, and Implications," *Brookings Papers on Economic Activity*, 34, pp. 207–98.
Greene, D. (2008), *Econometric Analysis*, Prentice Hall.
Greenspan, A. (1998), "Is there a new economy?," Remarks at the Haas Annual Business Faculty Research Dialogue, University of California, Berkeley, 4 September.
Greenwood, J., and B. Jovanovic (1990), "Financial Development, Growth, and the Distribution of Income," *Journal of Political Economy*, 98, pp. 1076–106.
Griffith, R., and R. Harrison (2004), "The Link between Product Market Reform and Macro-economic Performance," *European Commission Economic Papers*, no. 209.
Grossman, G., and E. Helpman (1989), "Product Development and International Trade," *Journal of Political Economy*, 97, pp.1261–83.
Guellec, D. (1999), *Economie de l'innovation*, La Découverte.
Gurley, J., and E. Shaw (1955), "Financial Aspects of Economic Development," *American Economic Review*, 45, pp. 515–38.
Hahn, F., and R. Matthews (1964), "The Theory of Economic Growth: A Survey," *Economic Journal*, 74, pp. 779–902.
Hamilton, K. (2006), *Where is the Wealth of Nations? Measuring Capital for the 21st Century*, The World Bank
Harrod, R. (1939), "An Essay in Dynamic Theory," *Economic Journal*, 49, pp. 14–33.
Hausmann, R., D. Rodrik, and A. Velasco (2008), "Growth Diagnostics," in N. Serra and J.Stiglitz (eds.), *The Washington Consensus Reconsidered: Towards a New Global Governance*, Oxford University Press.
High Level Group (2004), "Facing the Challenge: The Lisbon Strategy for Growth and Employment" (Kok Report), November, Brussels: European Commission.
Hotelling, H. (1929), "Stability in competition," *Economic Journal*, 39, pp. 41–57.

International Monetary Fund (2003), *World Economic Outlook*.
Iwata, S., M. Khan, and H. Murao (2003), "Sources of Economic Growth in East Asia: A Nonparametric Assessment," *IMF Staff Papers*, 50, pp.157–76.
Jones, C. (2005), "Growth and Ideas," in P. Aghion and S. Durlauf (eds.), *Handbook of Economic Growth*, Elsevier, Chapt. 16, pp. 1063–111.
Kanbur, R. (2000), "Income Distribution and Development," in A. Atkinson and F.Bourguignon (eds.), *Handbook of Income Distribution*, Elsevier, Chapter 13, pp. 791–841.
Kaufman, D., A. Kraay, and M. Mastruzzi (2008), "Governance Matters VII: Aggregate and Individual Governance Indicators, 1996–2007," World Bank Policy Research Working Paper no. 4654.
Kose, M.A., E. Prasad., K. Rogoff, and S.-J. Wei (2006), "Financial Globalization: A Reappraisal," IMF Working Paper no. 06/189, August.
Kremer, M. (1993), "Population Growth and Technological Change: 1,000,000 B.C. to 1990," *The Quarterly Journal of Economics*. August, pp. 681–716.
Krugman, P. (1991a), "Increasing Returns and Economic Geography," *Journal of Political Economy*, 99, pp. 483–99.
Krugman, P. (1991b), "History versus Expectations," *Quarterly Journal of Economics*, 106, pp. 651–67.
Krugman, P. (1991c), *Geography and Trade*, MIT Press.
Krugman, P. (1994a), "The Fall and Rise of Development Economics," in L. Rodwin and D.A. Schön (eds.), *Rethinking the Development Experience: Essays Provoked by the World of Albert Hirschman*, Washington D.C.: The Brookings Institution Press, pp. 39–58.
Krugman, P. (1994b), "The Myth of Asia's Miracle," *Foreign Affairs*, November–December, pp. 62–78.
Krugman, P. (1995), *Development, Geography, and Economic Theory*, MIT Press.
Kuhn, T. (1962), *The Structure of Scientific Revolutions*, University of Chicago Press.
Kuznets, S. (1955), "Economic Growth and Income Inequality," *American Economic Review*, 65, pp.1–29.
La Porta, R., F. Lopez-de-Silanes, A. Schleifer, and R. Vishny (1998). "Law and Finance," *Journal of Political Economy*, 106, pp. 1113–55.
La Porta, R., F. Lopez-de-Silanes, A. Schleifer, and R. Vishny (1999), "The Quality of Government," *Journal of Law, Economics and Organization*, 15, pp. 222–79.
Le Gallo, S., and S. Dall'erba (2006), "Evaluating the Temporal and Spatial Heterogeneity of the European Convergence Process, 1980–1999," *Journal of Regional Science*, 46, pp. 269–88.
Levine, R. (2005), "Finance and Growth: Theory and Evidence," in P. Aghion and S. Durlauf (eds.), *Handbook of Economic Growth*, Elsevier, Chapt. 12, pp. 865–934.
Lucas, R. (2004), "Industrial Revolution, Past and Future," *Annual Report 2003*, Federal Reserve Bank of Minneapolis (http://minneapolisfed.org/pubs/region/04-05/essay.cfm)
Maddison, A. (1997), "Causal Influences on Productivity Performance 1820–1992," *Journal of Productivity Analysis*, November, pp. 325–60.
Maddison, A. (2001), *The World Economy: A Millennial Perspective*, OECD Development Centre, Paris: Organization for Economic Cooperation and Development.
Maddison, A. (2007), *Contours of the World Economy 1–2030 AD*, Oxford University Press.

Malthus, T. (1798), *An Essay on the Principle of Population, as It Affects the Future Improvement of Society with remarks on the Speculations of Mr. Godwin, M. Condorcet, and Other Writers*, J. Johnson.

Mankiw, N.G., D. Romer, and D.N. Weil (1992), "A Contribution to the Empirics of Economic Growth," *The Quarterly Journal of Economics*, 107, 407–37.

Martin, P. (1999), "Public policies, regional inequalities and growth," *Journal of Public Economics*, 73, pp. 85–105.

McKinnon, R. (1973), *Money and Capital in Economic Development*. Washington, D.C. The Brookings Institution.

Meisel, N. and J. Ould Aoudia (2008), "Is 'Good Governance' a Good Development Strategy?", *Working Paper no. 58*, Agence française de développement.

Milanovic, B. (2005), *Worlds Apart: Measuring International and Global Inequality*, Princeton University Press.

Munnell, A. (1992), "Infrastructure, Investment and Economic Growth," *Journal of Economic Perspectives*, 6, pp. 189–98.

Murphy, K., A. Schleifer, and R. Vishny (1989), "Industrialization and the Big Push," *Journal of Political Economy*, 97, pp. 1003–26.

Nelson, R., and E. Phelps (1966), "Investment in Humans, Technological Diffusion, and Economic Growth," *American Economic Review: Papers and Proceedings* 51, pp. 69–75.

Nicoletti, G., and S. Scarpetta (2005), "Regulation and Economic Performance. Product Market Reforms and Productivity in the OECD," *OECD Economics Department Working Paper*, no. 460, Paris: Organisation for Economic Cooperation and Development.

North, D. (1990), *Institutions, Institutional Change and Economic Performance*, Cambridge University Press.

North, D. (1994), "Institutions Matter," revision of 28/03/94, available at http://129.3.20.41/econ-wp/eh/papers/9411/9411004.pdf.

North, D., J. Wallis, and B. Weingast (2009), *Violence and Social Orders. A Conceptual Framework for Interpreting Recorded Human History*, Cambridge University Press.

OECD (2003), *ICT and Economic Growth: Evidence from OECD Countries, Industries and Firms*, Paris: Organisation for Economic Cooperation and Development.

OECD (2004), *Economic Survey of the Euro Area*, Paris: Organisation for Economic Cooperation and Development.

OECD (2010), "The Impact of the Economic Crisis on Potential Output," *Working Party No. 1 on Macroeconomic and Structural Policy Analysis, ECO/CPE/WP1(2010)3*, Paris: Organisation for Economic Cooperation and Development.

Oliner, S., and D. Sichel (2002), "Information Technology and Productivity: Where Are We Now and Where Are We Going?," *Federal Reserve Bank of Atlanta Economic Review*, third quarter.

Pagano, M. (1993), "Financial Markets and Growth: An Overview," *European Economic Review*, 37, pp. 613–22.

Perée, É., and T. Välilä (2007), "A Primer on Public Investment in Europe, Old and New," European Investment Bank, *Economic and Financial Report*, no. 2007/01.

Philippon, T., and N. Véron (2008), "Financing Europe's Fast Movers," *Bruegel Policy Brief* no. 2008/01, Bruegel, Brussels.

Rajan, R. and L. Zingales (2003), *Saving Capitalism from the Capitalists*, Random House.

Ramey, G., and A. Ramey (1995), "Cross-Country Evidence on the Link between Volatitity and Growth," *American Economic Review*, 85, No. 5.
Ramsey, F. (1928), "A Mathematical Theory of Saving," *Economic Journal*, 28, pp. 543–59.
Ray, D. (2001), "What's New in Development Economics," *American Economist*, 44, pp. 3–16.
Rodriguez, F., and D. Rodrik (2001), "Trade Policy and Economic Growth: A Skeptic's Guide to the Cross-National Evidence," in Bernanke, B. and K. Rogoff (eds.), *Macroeconomics Annual 2000*, MIT Press for NBER.
Rodrik, D. (2004), "Getting institutions right," CESifo DICE Report, No 2.
Rodrik, D. (2005), "Growth strategies," in Aghion, P., and S. Durlauf (eds.), *Handbook of Economic Growth*, Elsevier, Chapt. 14, pp. 967–1014.
Rodrik, D. (2006), "Goodbye Washington Consensus, Hello Washington Confusion? A Review of the World Bank's *Economic Growth in the 1990s: Learning from a Decade of Reform*," *Journal of Economic Literature*, 64, pp. 973–87.
Romer, P (1986), "Increasing Returns and Long-run Growth," *Journal of Political Economy*, 94, pp. 1002–36.
Romer, P. (1990), "Endogenous Technological Change," *Journal of Political Economy*, 98, pp.71–102.
Rosenstein-Rodan, P. (1943), "Problems of Industrialization of Eastern and Southeastern Europe," *Economic Journal*, 53, pp. 202–11.
Sala-ì-Martin, X. and M. Pinkovskiy (2010), "African Poverty is Falling ... Much Faster than You Think!" *NBER Working Paper* no. 15775.
Sapir, A. et al. (2004), *An Agenda for a Growing Europe*, Report to the President of the European Commission, Oxford University Press.
Schiantarelli, F. (2005), "Product Market Regulation and Macroeconomic Performance: A Review of Cross Country Evidence," IZA Discussion Paper no. 1791.
Schumpeter, J. (1911), *The Theory of Economic Development*, Harvard University Press.
Schumpeter, J. (1942), *Capitalism, Socialism and Democracy*, Harper & Row (reprinted 1976 George Allen & Unwin).
Sen, A. (1999), "The Possibility of Social Choice" (Nobel Lecture), *American Economic Review*, 89, July.
Sen, A. (2000), "A Decade of Human Development," *Journal of Human Development*, 1, pp. 17–23.
Shaw, E. (1973), *Financial Deepening in Economic Development*, Oxford University Press.
Sianesi, B., and J. Van Reenen (2002), "The Returns to Education: A Review of the Empirical Macro-Economic Literature," Institute for Fiscal Studies, WP 02/05.
Solow, R. (1956), "A Contribution to the Theory of Economic Growth," *Quarterly Newspaper of Economics*, 70, pp. 65–94.
Stiglitz, J., A. Sen, and J.P. Fitoussi (2009), *Report by the Commission on the Measurement of Economic Performance and Social Progress*, Paris.
Swan, T. (1956), "Economic Growth and Capital Accumulation," *Economic Record*, 32, pp. 334–61.
Timmer, M., G. Ypma, and B. van Ark (2003), "IT in the European Union: Driving Productivity Divergence?," *Groningen Growth and Development Centre Research Memorandum GD-67*, Groningen: University of Groningen.

Tirole, J. (2003), "Protection de la propriété intellectuelle : une introduction et quelques pistes de réflexion," in *Propriété intellectuelle*, rapport au Conseil d'Analyse Economique 41, Paris: La Documentation Française, pp. 9–48.

United Nations Development Program (2007), *Human Development Report 2007/2008.*

Van Ark, B., and E. Bartelsman (2004), "Fostering Excellence: Challenges for Productivity Growth in Europe," Background document for the Informal Competitiveness Council Maastricht, 1–3 July.

Van Ark, B., and R. Inklaar (2005), *Catching Up or Getting Stuck? Europe's Troubles to Exploit ICT's Productivity Potential*, Research Memorandum GD 79, Groningen: Groningen Growth and Development Centre, September.

Wallerstein, I. (1979), *The Capitalist World-Economy*, Cambridge University Press.

Wilson, D. (2005), "The Rise and Spread of State R&D Tax Credits," *Federal Reserve Bank of San Francisco Economic Letter*, no. 2005–26.

Yang, S., and E. Brynjolfsson (2001), "Intangible Assets and Growth Accounting: Evidence from Computer Investments," Working Paper, Massachusetts Institute of Technology, May, available as Paper 136 at the Center for eBusiness@MIT (http://ebusiness.mit.edu).

Young, A. (1928), "Increasing Returns and Economic Progress," *Economic Newspaper*, 38, pp. 527–42.

Young, A. (1992), "A Tale of Two Cities: Factor Accumulation and Technical Change in Hong Kong and Singapore," in Blanchard, O., and Fischer, S., *NBER Annual Macroeconomics,* MIT Press, pp. 13–54.

Young, A. (1995), "The Tyranny of Numbers: Confronting the Statistical Realities of the East Asian Growth Experiment," *The Quarterly Journal of Economics*, 110, pp. 641–80.

7

Tax Policy

7.1 Issues
- 7.1.1 What is taxation about?
- 7.1.2 Typologies of tax systems
- 7.1.3 Taxation in developing economies
- 7.1.4 Redistribution versus efficiency

7.2 Theories
- 7.2.1 Tax incidence on a specific market
- 7.2.2 Social losses and distortions related to taxation
- 7.2.3 Tax incidence in general equilibrium
- 7.2.4 Effectiveness versus equity: Optimum taxation
- 7.2.5 Corrective taxation
- 7.2.6 Taxation in open economies

7.3 Policies
- 7.3.1 Distributing the tax burden efficiently
- 7.3.2 Distributing the tax burden equitably
- 7.3.3 Correcting market failures
- 7.3.4 Tax cooperation

References

> Taxation without representation is tyranny.
>
> James Otis (1725–83), US lawyer and politician at the time of the American Revolution

> A common contribution is essential for the maintenance of the public forces and for the cost of administration. This should be equitably distributed among all the citizens in proportion to their means.
>
> French Declaration of the Rights of Man and of the Citizen, Article 13, 1789

Governments seldom charge for their services. Except in specific areas such as museums, swimming pools, or universities, these services are generally provided free of charge. Think about scientific research, defense, or diplomacy.

Those are public goods (see chapter 2); it is therefore impossible to identify precisely which citizens benefit from them and to have them pay for their consumption. Such public services[1] have to be financed through *taxation**, i.e., compulsory contributions by households or corporations.

Not all tax-financed government services are public goods. Think of schooling: It would be possible for the government to charge for its provision, but the common practice is to finance it at least partly through taxation. It is by their vote that citizens can express preferences regarding the level and quality of public services provision and the corresponding level of taxes.

The problem is that taxes generally distort relative prices. For instance, personal income tax is paid by households on their income, which generally accrues to their labor. This increases the relative price of labor as compared to leisure and may therefore change the labor supply. Such distortions may impact welfare and GDP growth. Consequently, there is a trade-off between the provision of public goods (which in many cases, such as education, security, or infrastructure building, is expected to have a positive impact on welfare and growth) and the desire to reduce taxation in order to limit price distortions.

In some cases, however, taxation reduces distortions and contributes to efficiency. The best example is that of pollution: Absent taxation, pollution is produced in excess because there is no price attached to it. So-called "green" taxes that put a price on clean air or water contribute to a cleaner environment and therefore improve welfare.

From a pure efficiency standpoint, public services should be financed through *lump-sum taxes**, i.e., taxes that are levied in equal amounts on every citizen independently of their activity, consumption, or income, because such taxes do not distort work, saving, and consumption decisions. However, such taxation (experimented with by Margaret Thatcher's government in 1979 through the so-called *poll tax**) is questionable from an equity standpoint, since the poor pay relatively more as a percentage of their income than the wealthy. In order for the burden of taxation to be distributed in an equitable way, and even more when income redistribution is a policy objective, taxes have come to be proportional or more than proportional to income, which inevitably introduces economic distortions.

Because it is at the heart of the efficiency–equity trade-off, and because it is the simplest way to redistribute wealth among citizens, tax policy has always and everywhere been hotly politicized, often at the cost of overlooking essential economic considerations. Tax policy is a matter for political decision and is in all democracies a prerogative of parliaments,[2] Its consequences, however,

1. In this chapter, we speak of *government services** to designate all goods produced, or services provided, by governments (or in some cases public enterprises) whatever the justification (or the lack of it) for their public character. We reserve the expression *public goods* to goods and services whose consumption is neither excludable nor rivalrous, as defined in chapter 2.

2. A fine line has to be drawn between the prerogatives of the government and of parliament. For instance, the French Parliament determines "the rules concerning the base, rates and methods of

are often visible in the long term only. As we shall show in this chapter, economic analysis can greatly contribute to the design of efficient taxation systems. The role of economists is to alert policymakers to the economic and social consequences of alternative tax-policy choices and to help determine how social preferences can best be served by adequate tax instruments.

7.1 Issues

7.1.1 What is taxation about?

*Tax policy** consists in setting, within the annual budget, the *rate** and the *base** of each tax. For instance, the government can decide to increase the consumption tax by one percentage point (a decision on the rate) or to exempt some items from this tax (a decision on the base). *Tax revenues** depend on the combination of rates and bases. It is possible to maintain a given level of revenues through simultaneously cutting the rate and broadening the base (this has been the general trend recently as we shall see later on). In most cases, choices on the rate and on the base interact: For example, taxes levied on pollution aim at reducing pollution. More generally, tax revenues cannot be precisely determined *ex ante* because tax bases themselves depend on economic activity: For instance, for a given definition of rates and bases, a downturn in, say, consumption spending will automatically reduce the income accruing to consumption taxes.

a) Why taxation?

Tax policy is at the crossroads between the three functions of economic policy identified by Richard Musgrave and listed in chapter 1:

- *Allocation.* Taxation affects relative prices between goods and services, labor and leisure, labor and capital, etc. In so doing, it creates *price distortions**: For example, a tax on imports increases the price of foreign goods relative to the domestic ones. Only lump-sum taxes are nondistortionary. Such taxes exist in practice (e.g., hotel or airport taxes in some countries) but they are few. In a perfect market economy, i.e., where relative price adjustments maintain an optimum allocation of resources (cf. the first theorem of welfare discussed in chapter 1), taxes would typically be detrimental to economic efficiency. However, the presence of market imperfections modifies this diagnosis. For instance, taxation makes it possible to correct externalities such as air pollution: Without taxation, industries would pollute more than what is socially optimal. Taxation also makes it possible to finance public goods that

collection of taxes of all types" (Art. 34 of the French Constitution of 1958), which implies that the right of the government to set tax rates by decree is strictly limited.

would not be spontaneously produced by the markets. Lastly, it can play a "*paternalistic**" role by protecting private agents from their own errors. Taxes on alcohol and cigarettes are examples of such paternalist taxes, sometimes referred to as *sin taxes**.

- *Distribution.* Income taxation modifies the distribution of income between rich and poor, between families and single individuals, or between generations. Capital taxation (either at the firm or at the household level) and social insurance[3] contributions (levied on labor) affect the relative income shares of capital and labor. Distributional effects can be involuntary but are also sought by governments when the market equilibrium is regarded as contrary to equity. Since the French revolution of 1789, it has been increasingly admitted that taxation should be either *proportional** to income, or *progressive** (more than proportional, meaning that the rich pay relatively more in proportion of their income), but not *regressive** (less than proportional, meaning that the rich pay relatively less) as was often the case previously.[4]
- *Stabilization*: As shown in chapter 3, a lower tax burden during a cyclical downturn helps sustain the demand for goods and services, and conversely, higher taxes during a boom slow down demand, alleviating upward pressures on prices. Automatic stabilization, i.e., stabilization performed at constant tax rates through the endogenous adjustment of tax bases, is usually distinguished from discretionary stabilization through decisions to change tax rates and bases counter-cyclically. However, constraints on public finances as well as political pressures can lead the government to raise tax rates during economic downturns, and reduce them when the economy is booming. Such procyclical policies accentuate, rather than dampen, business cycles. The stabilization role of tax policy, already discussed in chapter 3, will not be further addressed in this chapter.

These three functions of taxation are closely interconnected and often give rise to trade-offs. For instance, automatic stabilizers (stabilization function) are larger in countries that have higher levels of taxation designed to redistribute more income among the residents (redistribution function) or to produce more government services (allocation). Typically, automatic stabilizers are larger in the euro area than in the US. A progressive income tax is able to reduce income inequalities across households (redistribution), but it can also reduce the incentive to work and therefore affect economic

3. In Europe, contributions to health, old-age, and unemployment insurance are often called *social security contributions**, but in the US only the federal old-age insurance program is called *social security**. To avoid confusion, we speak of *social insurance contributions** rather than social security contributions.

4. In the Middle Ages, serfs would pay contributions in money and in kind to finance the expenditures of the Lords. The French Revolution introduced the notion of proportional taxes, but progressive taxation only appeared at the beginning of the twentieth century.

efficiency (allocation). A tax on cigarettes reduces diseases and increases tax revenues (allocation) but generally has regressive effects, meaning that the poor pay relatively more.[5]

b) How much?

During previous historical periods, taxation was almost exclusively determined by wars: In peacetime, taxes would represent a very low share of national income, whereas kings and emperors would raise taxes to finance wars, whatever the social consequences. The beginning of the twentieth century still followed this pattern, with taxation representing less than 10% of GDP before World War I, and reaching or even exceeding 50% of GDP for some belligerents during each of the two world wars. In the US, the top marginal income tax rate reached 77% in 1918 and 94% in 1945 and the tax base was greatly extended, whereas only 2% of the population paid this tax in 1915.[6]

Between World Wars I and II, total taxation decreased as a percentage of GDP. However, this decline was counter-balanced by the birth of the *welfare state**, i.e., the system of social protection. Compulsory health and old-age insurance had started to develop in Germany in the late nineteenth century under Chancellor Bismarck and had extended to other European countries (but not to the US). Coverage was extended in the twentieth century. In the US, the New Deal of Franklin Delano Roosevelt introduced federal social programs that contributed to a significant rise in federal taxes (see chapter 3). The trend accelerated after World War II with the introduction of comprehensive social insurance regimes covering unemployment, aging, health, and poverty risks. These systems involved a steady increase in *total taxation** (i.e., in the aggregate burden of all taxes).

In the 1980s, a divide emerged between, on the one hand, the further development of social protection in continental Europe, and on the other hand, a rollback on welfare development in the US and several other English-speaking countries. Consistently, total taxation continued to rise in continental Europe, while it stabilized around a constant level in the US and decreased in the UK. In the late 1990s and in the 2000s, however, governments in continental Europe made substantial efforts to stabilize or even curb total taxation, whereas a renewed emphasis on government services led total taxation to rise again in the UK, where it reached a level close to that of continental Europe. Canada also dramatically curbed total taxation in the 1990s and 2000s, whereas Japanese taxation stabilized at a level close to that of the US (figure 7.1).

Within the EU, total taxation varied from 29% of GDP in Romania to 49% in Denmark in 2007 (figure 7.2). The so-called *New Member States** which

5. In 2000, according to Godefroy (2003), taxes on cigarettes represented 5.25% of income in the first decile of French household income and 0.48% for the last decile, against an average of 1.4%.
6. See Salanié (2003).

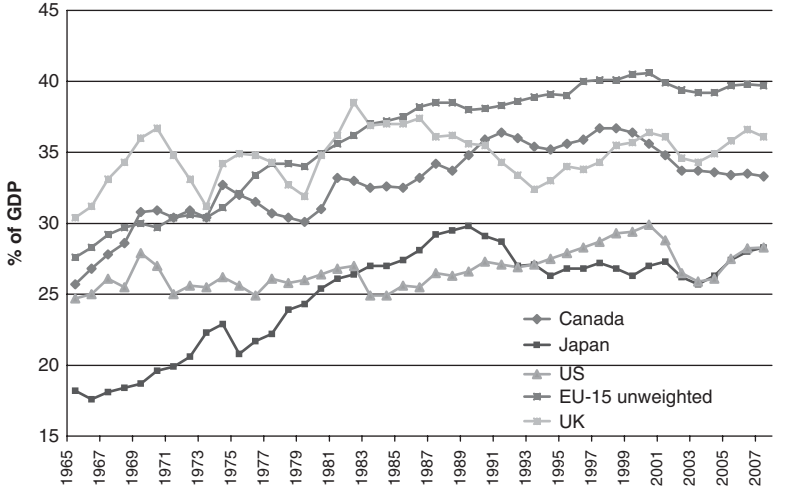


Figure 7.1 Total tax revenues in selected countries/areas, in % of GDP, 1965–2007.
Source: OECD, *Revenue Statistics*, 2009.

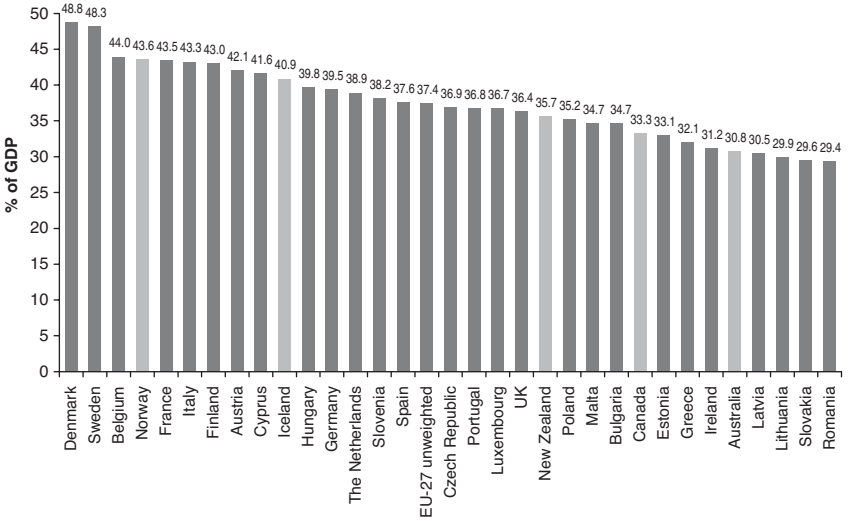


Figure 7.2 Total tax revenues in selected countries, in 2007 (% of GDP).
Source: European Commission, *Taxation Trends in the European Union*, 2009; OECD, *Revenue Statistics*, 2009.

joined the EU in 2004 and 2007,[7] generally display lower levels of taxation. At the other end of the spectrum, Scandinavian economies display the highest levels of total taxation, together with Belgium, France, and Italy. Germany used to be part of this high-tax group, especially in the aftermath of reunification, but the level of taxation declined significantly in the early 2000s.

The wide dispersion of total taxation rates across countries of similar development levels, even within the EU, points to significant differences in national preferences for the provision of government services: Taxation is high in Scandinavian countries where a large range of educational, health, and social services are available for free and financed by taxes, and it is low in the US where similar services are provided by the private sector. An indication that this difference can be ascribed to preferences is that differences in total taxation levels have not narrowed over the last decades in spite of the much tighter integration of product and capital markets. Thus, the widespread expectation that globalization would force convergence does not seem to be supported by the data. We will return to the issue when examining the consequences of openness for tax policies.

7.1.2 Typologies of tax systems

Taxes can be classified along three dimensions depending on: (i) Who collects them, (ii) how they are collected, and (iii) who pays them.

a) Who collects taxes?

Taxes can be levied by the central government, state governments (especially in federal countries, e.g., *Länder* in Germany, *cantons* in Switzerland, *provinces* in Argentina), local governments, and social insurance administrations. However, the administration that levies the tax may not be the one that decides on it or benefits from it. For instance, local taxes can be levied by the central tax administration on behalf of local authorities.

Figure 7.3 illustrates the diversity of tax structures in the European Union.[8] In Germany, Belgium, and Scandinavian countries, states, linguistic communities and/or local authorities collect an important share of the public revenues. In Sweden, for example, two-thirds of funds for public expenditures on health are collected at the local level. In marked contrast, taxation is highly centralized in countries like Greece, Ireland, The Netherlands, the UK, and the Czech Republic. France has the highest share of social insurance contributions.

The distribution of taxes between the central and local governments raises issues of *tax autonomy** and *tax competition**, in the context of Oates' equivalence discussed in chapter 2. On the one hand, it is desirable that

7. These are: Bulgaria, Cyprus, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Romania, Slovakia, and Slovenia.
8. Contributions to the EU budget are not included in the graph.

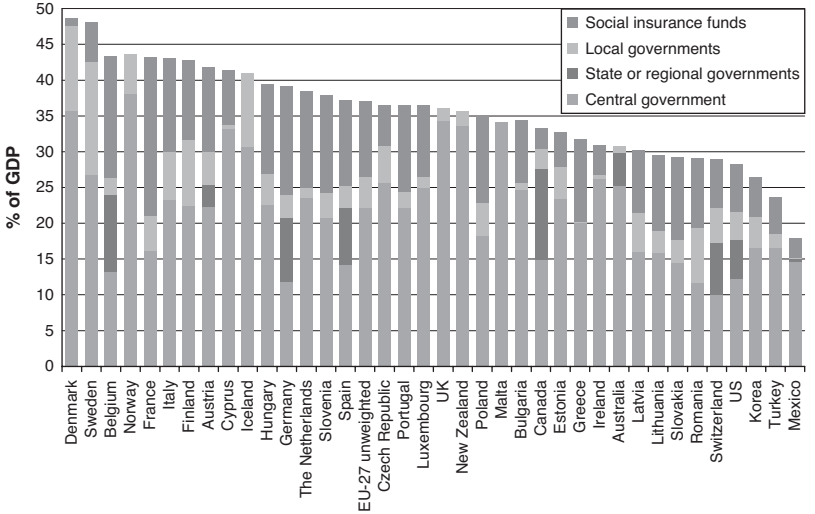


Figure 7.3 Structure of taxation according to the administration of perception, in 2007 (% of GDP).
Source: European Commission, *Taxation Trends in the European Union*, 2009; OECD, *Revenue Statistics*, 2009, and OECD *Revenue Statistics*, 2009.

taxes be raised at the local level to make local governments financially responsible and to allow them to develop policies that are consistent with local preferences (tax autonomy). On the other hand, there is a risk that autonomy would allow wealthy localities to become even wealthier because they are able to attract more individuals and companies through cutting taxes (tax competition), whereas poor localities would need to raise tax rates because the tax base is limited and because they have higher social expenditures. Central governments therefore frequently organize redistribution across localities. The level of this redistribution is a contentious issue since it affects the trade-off between efficiency (of local public choices) and equity (between localities).

b) How are taxes collected?

Another classification of taxes relies on the way they are collected.

A *direct tax** is a tax levied on income (or wealth) whatever the use of this income (or wealth). Direct taxes include:

- for households, the *personal income tax** (a tax on labor and capital income that can be paid directly by the households or levied by the employers), *inheritance taxes**, *property taxes**,[9] and *wealth taxes**;

9. Property taxes are based on the market value of land and housing held by households, with a tax rate generally decided at the local level.

- for companies, the *corporate income tax**;[10] and local business taxes such as the German *Gewerbesteuer*, the French *contribution économique territoriale* or the Italian IRAP.

In contrast, an *indirect tax** is levied on the use of income, mainly on consumption. Typical examples include taxes levied on imports of goods and services (import duties), the US sales tax and the European *value-added tax** (*VAT**),[11] both of which are borne by consumers when they buy a good or a service. *Excise taxes**, i.e., taxes on miscellaneous products such as cigarettes or alcohol, are other examples of indirect taxes. Finally, environmental taxes (including energy, transport, and pollution taxes) are also indirect taxes.

The third category of taxes covers *social-insurance contributions**, that are paid both by employers and by employees on the basis of the wage bill. Although they can be considered as direct taxes, social insurance contributions are generally treated separately, which makes sense in "Bismarckian" countries where there is, in principle, a direct link between individual contributions and benefits (see infra).

Figure 7.4 shows the 2007 structure of taxation in a number of countries according to this second classification. In Scandinavian countries, the UK, Ireland, and Belgium, direct taxes are predominant, representing 40% or more of total taxation (61%, in the case of Denmark). Conversely, in France, Germany, the Czech Republic, and Slovakia, social contributions represent around 40% of total taxation. Finally, the new EU member states generally rely heavily on indirect taxation.

The disparity of tax systems illustrated by figure 7.4 embodies a differentiation between *Bismarckian systems** (Germany, Austria, France, Sweden, The Netherlands) and *Beveridgian systems** (the UK, Denmark, Ireland). In the first ones, inspired by the scheme introduced by German Chancellor Bismarck in the 1880s, social insurance benefits are treated as deferred wages; they are therefore financed primarily by social contributions based on wages; each employee knows that what s/he will receive when unemployed or retired, will be proportional to his or her contribution. In the second system, introduced in the UK after William Beveridge's 1941 report, social benefits are viewed as public transfers whose objective is to ensure that the most deprived receive a minimum level of income; they are financed primarily through taxes and there is little link, at the individual level, between contributions and benefits. With time, the contrast between the two schemes has tended to fade

10. The corporate income tax is levied on companies' profits, after tax allowances (such as accelerated capital depreciation, or R&D allowances) have been deducted.

11. VAT is not a tax on value added, but a tax on final consumption. It is charged on the purchase of goods and services as a percentage of the value of these goods. VAT is then transferred to the tax administration by the seller of the goods. VAT paid on investment goods and on intermediate consumption is recovered by the firms. The fact that intermediate consumption is not taxed ensures that the same good is not taxed several times (as an intermediate consumption, then as a final consumption) and explains why this tax is named value-added tax, see box 7.11.

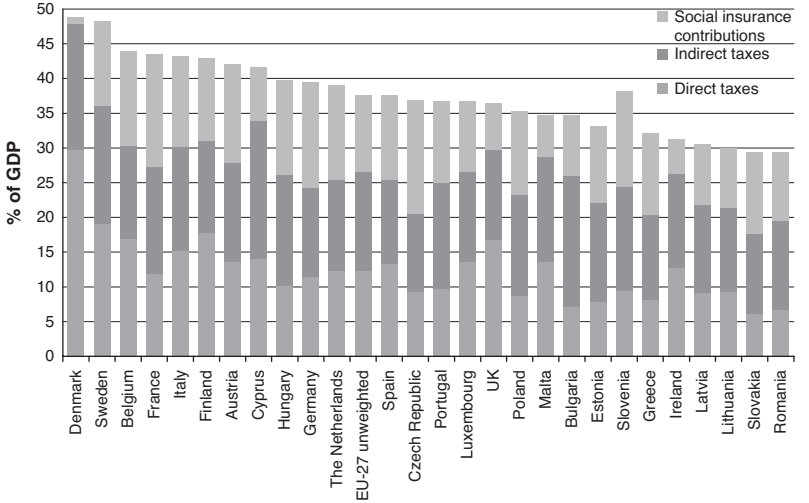


Figure 7.4 Structure of taxation in EU countries, in 2007 (% of GDP).
Source: European Commission *Taxation Trends in the European Union*, 2009; OECD, *Revenue Statistics*, 2009.

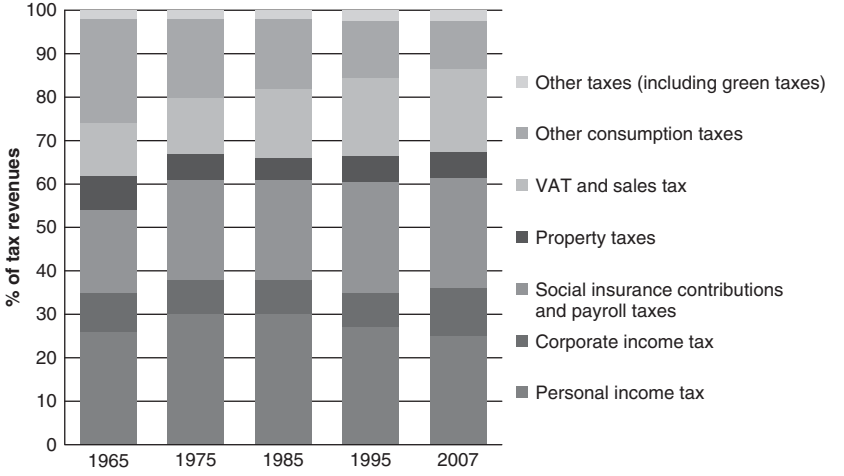


Figure 7.5 Revenue share of taxes in the OECD area (unweighted average).
Source: OECD, *Revenue Statistics*, 2009.

away: Beveridgian systems have introduced some insurance schemes, whereas Bismarckian ones have been complemented with redistributive transfers. In some countries, there has also been a shift from social contributions to indirect taxes, notably environmental taxes (Denmark, Sweden), income taxes (France) and VAT (Denmark, Germany).

Since the mid-1960s, the share of indirect taxes in total tax revenues has tended to decline in advanced countries, but this aggregate evolution results from the opposite trends of declining excise taxes and rising VAT revenues (figure 7.5). In fact, in a world of mobile capital and labor income bases, VAT has been increasingly regarded as an efficient, nondistortionary way to raise revenues. Another evolution since the 1980s has been a fall in the share of personal income tax revenues (over total tax revenues) from almost 30% in 1985 to 25% in 2007. Governments in many countries have reduced politically sensitive personal income taxes while increasing both social insurance contributions and VAT rates. Section 7.2 will help us discuss the rationale for these choices.

Although popular, the distinction between direct taxes and indirect taxes is hardly relevant for economic analysis. An important insight from the economics of taxation is that the agent that ultimately bears the tax burden is often not the one who pays the tax bill to the tax authority. Accordingly, the distinction between direct and indirect taxes matters more for the practical organization of tax collection than for the analysis of the effects of taxation (cf. section 7.2).

An "economic" (rather than legal) definition of direct taxes is given by Tony Atkinson (1977): Direct taxes are those that can be personalized, i.e., adapted to the taxpayer's characteristics. For instance, personal income tax depends on the household's characteristics and on the nature of income received (labor income, capital income, pensions, or social transfers). Similarly, corporate income tax depends on taxable profit that takes into account recent investment or R&D expenditures; in many countries, the tax rate is also different depending on the size of the company or on the use of profit (whether it is distributed as dividends or reinvested in the company). In contrast, indirect taxes are levied on anonymous transactions, any taxpayer thus faces the same tax rate.From this economic definition of direct and indirect taxes, it follows that only the former can be used for redistribution purposes. Indeed, direct taxes are largely used to redistribute income from rich to poor households, from single individuals to families, or from large companies to SMEs. The income tax can also allow *negative taxation**, which makes it possible to extend redistribution at the lower end of the income scale (cf. box 7.1)

Box 7.1 Negative Taxation

The idea of a negative tax was introduced in 1946 by George Stigler, and taken up again in 1962 by Milton Friedman, both free-market economists. Its principle is to extend the progressiveness of the personal income tax in such a way that the tax becomes negative (rather than just zero) below a certain income threshold, which makes it possible to improve the income of poor households (cf. figure B7.1.1).

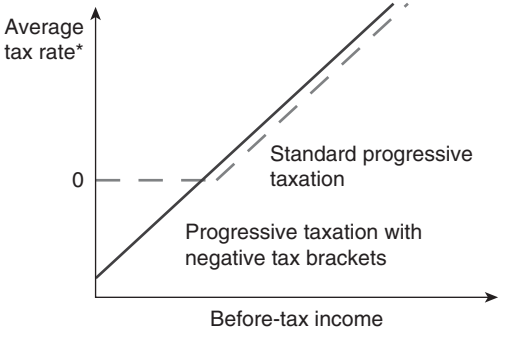


Figure B7.1.1 Negative taxation.
Note: *Tax paid as a proportion of income.

In the US, the idea was successively endorsed by both a democratic President, Lyndon Johnson (1963–69), and a Republican, Richard Nixon (1969–74). A form of negative tax finally emerged in 1975 with the *Earned Income Tax Credit** (EITC). The EITC nevertheless differs from the negative tax in that it is not universal; rather, it is limited to those families where at least one adult is employed—thus serving as a work incentive for welfare recipients. The EITC remained of secondary importance until the mid-1980s, when a series of reforms (in 1986, 1990, 1993, 2001) extended it. The Clinton administration (1993–2001) used it as a key instrument for fighting poverty at work. In the mid-2000s, the federal scheme[a] provided up to $4500 per family as supplementary income and benefited around 16% of households. It was found to lower child poverty by one-fifth and to have a strong positive impact on labor market participation by single-parent households. It had no effect on hours worked by those already employed and had a negative impact on participation for two-parent households, since the second parent would lose the benefit from the EITC when taking a job (Holt, 2006).

In 1999, Tony Blair introduced in the UK the *Working Family Tax Credit** (WFTC), a form of negative tax targeted at families with children where at least one adult was working. A minimum wage was simultaneously introduced in order for the WFTC not to be de facto transferred to employers through lower wages. The WFTC was more generous and more targeted (less than 10% of households) than the EITC. It was replaced in 2003 by the *Working Tax Credit** and the *Child Tax Credit**.

France also introduced in 2001 a tax credit system for families whose labor income lies between 0.3 times and 1.4 times the full-time minimum wage (up to 2.1 times the minimum wage in the case of single-worker families). This mechanism, called the *Employment bonus* ("*prime pour l'emploi*"*), is broad-based (in 2007, nine million persons benefitted) but not generous (an average of 480 euros a year per beneficiary in 2007),

even if, unlike the UK system, it can be topped up with other social insurance benefits (housing assistance, in particular). Its broad base limits its work-incentive impact (Cahuc, 2002). In 2008, the employment bonus was reformed with the introduction of the *Active Solidarity Income* ("*Revenu de solidarité active*" or *RSA*), a targeted benefit that can be combined with working income for low-paid workers.

Other countries that have introduced negative taxation schemes include Canada, Ireland, New Zealand, Finland, Belgium, The Netherlands, and Denmark.

[a]There were in addition 15 state-specific supplementary EITC schemes.

Indirect taxes are devoted to allocation functions, which consist both in financing the provision of public goods and in correcting market distortions. Note that these two objectives are largely incompatible, because what is aimed at is a stable tax base in the first case and a shrinking one in the second. This calls for using distinct instruments: On the one hand, a broad, inelastic base from which revenues can be raised without too many distortions; on the other, an elastic tax base to which a high tax rate can be applied.

c) Who pays?

Economists are generally reluctant to classify taxes according to the person who administratively pays the tax and makes the transfer to the tax administration—the *taxpayer**. For instance, they are not at ease with adding up employers' social contributions and corporate income tax, on the grounds that both are paid by corporations. Similarly, they prefer not to aggregate personal income taxes raised on labor income and those raised on capital income. They prefer to attach each tax to its tax base. Accordingly, a third classification of taxes distinguishes three categories: Labor, capital, and consumption. For instance, labor taxation covers social insurance contributions paid both by employers and employees, and lumps them together with the part of personal income taxes paid on labor income.[12]

Figure 7.6 shows the structure of EU countries' taxation systems across these three tax bases: Labor, capital, and consumption. The figure shows that consumption taxes represent roughly the same proportion of GDP across EU countries. Taxes on capital are lower in the new member states than in the "old" ones. Finally, figure 7.6 highlights that taxes on labor account for the bulk of between-country differences in total taxation: The countries with the highest total tax burden are also those where labor taxation is the heaviest.

12. In some countries, personal income taxes on labor income are raised by the employers and transferred by them to the tax administration. Although the employer then pays the entire package of labor taxes, these taxes can in fact bear on labor supplied by the workers, depending on the structure of the labor market (see section 7.2).

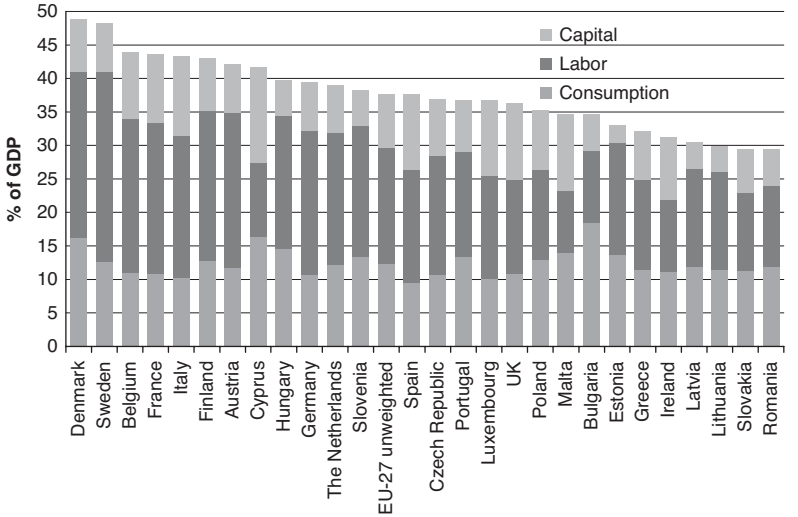


Figure 7.6 Tax structure according to tax bases in EU countries, 2007 (% of GDP).
Source: European Commission, *Taxation Trends in the European Union*, 2009.

7.1.3 Taxation in developing economies

Developing countries generally display a lower level of total taxation than advanced economies. In West Africa, for instance, total taxation varies from 11% of GDP in Guinea-Bissau to 20% in Senegal (2007 figures), against 30% to 50% of GDP in advanced economies. This is in part because taxation in developing countries encounters difficulties arising from low institutional development, corruption, and the large size of the informal sector. Those who hold political power and control natural resources often have the ability to escape taxation.[13] Furthermore, the demand for public services, like education, health, and infrastructure, increases with the level of income, thereby giving rise to higher public spending and taxation. While successful development rests on the ability to provide these public services, developing countries' governments typically lack the ability to collect adequate tax resources.

In this context, tax administrations retain an important discretionary power in the determination of tax bills and in the settlement of disputes. This especially affects direct taxes, in particular personal income tax, which therefore barely exists in developing countries. Social insurance contributions are also very limited, which reflects the low development of social protection systems, but also of wage-earning. Correspondingly, indirect taxes, especially tariffs, often play a central role in tax collection (see figure 7.7). This pattern

13. See Fjesldstad and Rakner (2003).

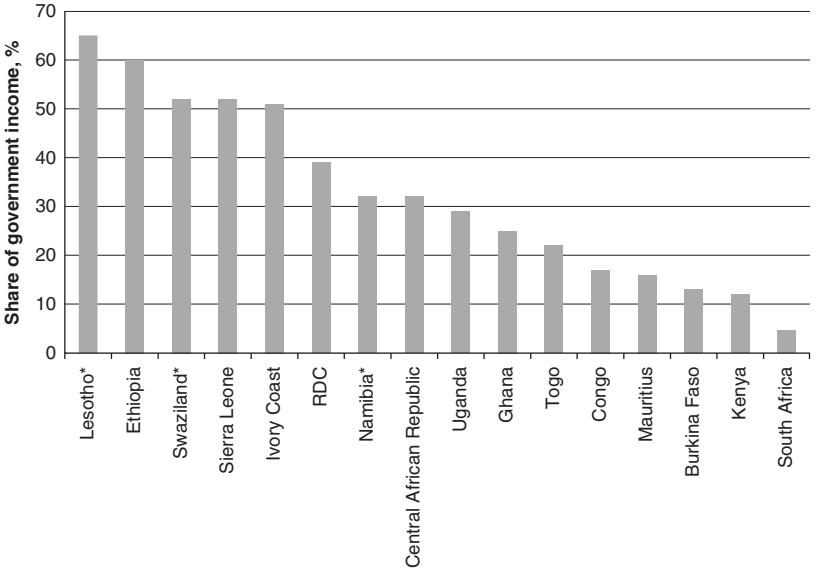


Figure 7.7 Share of duties in government income of selected, sub-Saharan African economies, in 2007 or closest (in %).
Source: IMF, Government Finance Statistics, December 2009.

Note: *Tariff revenues in Namibia and Swaziland partly result from South Africa sharing its tariff revenues with smaller members of SACU, the custom union among South Africa, Botswana, Lesotho, Namibia, and Swaziland.

partly explains the reservations expressed by developing countries, especially the least advanced ones, against tariff reductions negotiated within the framework of the World Trade Organization multilateral trade negotiations.[14]

7.1.4 Redistribution versus efficiency

As already mentioned, taxation plays a central role along two dimensions of public intervention: Redistribution and allocation. Generally, this involves a trade-off between redistribution and efficiency, as more redistribution requires more taxes that in turn are the source of additional distortions.

a) Redistribution

Taxes are often assessed according to their ability to reduce primary income inequalities. The redistributive impact of taxation cannot however be evaluated independently from that of *social transfers**, i.e., the benefits accruing

14. Emran and Stiglitz (2005) show that, when the informal economy is accounted for, substituting tariffs for VAT may be welfare decreasing.

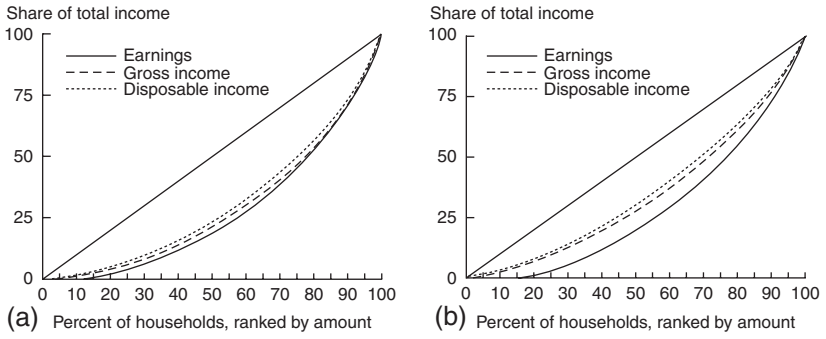


Figure 7.8 The Lorenz curve in the US and in Finland. a) US, b) Finland. Source: De Nardi et al. (2000), based on data from the Luxembourg Income Study database.

Notes: Earnings = gross wage, salary income, and farm and nonfarm self-employment income; gross income = earnings + cash property income (cash interest, rents, dividends, and annuities) + social and private transfers to households; disposable income = gross income − taxes paid by households. Data for 1994 (USA) or 1995 (Finland).

to households. Redistribution occurs mainly through the combination of the income tax, which is progressive, and of those social transfers that are conditional on resources, whether these transfers are pecuniary or in kind (education, health care). As presented in chapter 1, the extent of redistribution through transfers and taxes can be assessed by comparing the Lorenz curve based on market income to that based on disposable income (i.e., after tax and transfers). The latter is expected to be closer to the 45-degree line than the former. Figure 7.8 contrasts the Lorenz curves for the US against those for Finland. Although primary earning inequalities appear relatively similar in both countries, disposable income is less unequal in Finland, than in the US, thanks to net taxes and especially social transfers.

The degree of redistribution across income deciles depends on the *average tax rate**, i.e., the ratio between tax payments and income across income deciles: If the average tax rate increases with income, the tax system is deemed *progressive*; if it declines, it is said to be *regressive*; finally, if it is stable, the tax system is simply called neutral.

Redistribution does not occur solely across income levels (vertical redistribution) but also between categories of households, for instance between single persons and families, or two-parent and single-parent families. This horizontal redistribution aims at correcting income inequalities *per consumption unit**.[15]

15. To compare income across households, statisticians take into account the number of persons within a household as well as their relative consumption levels. According to an OECD definition, the first adult aged 18 and over represents one consumption unit, each subsequent adult aged 18

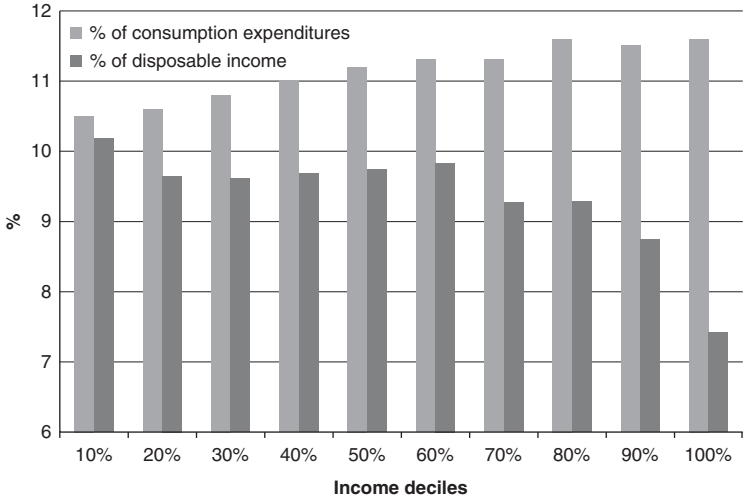


Figure 7.9 VAT as a regressive tax: VAT paid by decile of income in France. Source: French Ministry of Economy and Finance (2007), based on the 2001 income survey.

In the UK, both types of redistribution are combined since horizontal redistribution is targeted to the poorest families (see box 7.1).

VAT is generally seen as a regressive tax, because it is proportional to consumption, and poor households tend to save less and consume a higher share of their income than richer ones. Reduced VAT rates for basic-needs goods such as food and transportation and VAT exemption for rents, health services, insurance, and bank services partly compensate for this inequality. Still, as shown in figure 7.9 for the case of France, the percentage of disposable income devoted to VAT is a decreasing function of disposable income, which clearly makes VAT a regressive tax.[16]

Lastly, the redistributive effect of social contributions is limited. Social insurance contributions perform little redistribution between income levels because they are basically proportional to income—unless they are capped for higher incomes, in which case they are regressive. Turning to benefits, means-tested family allowances are redistributive, but pensions are roughly proportional to an individual's past income, and spending on health care tends to increase with income. This lack of redistribution through social contributions

and over represents 0.7 consumption units and each person aged under 18 accounts for 0.5 consumption units. A family composed of two adults and two children under 18 therefore represents 2.7 consumption units.

16. This analysis neglects the fact that savings are nothing other than a deferred consumption which will be taxed at a later date at the same VAT rates, in addition to taxes on capital income. Accounting for such delayed VAT payments, VAT no longer appears as a regressive tax.

is no accident, as Bismarckian systems have been designed to perform an insurance rather than a redistribution role. Benefits are considered as deferred incomes. This is less true in Beveridgian systems, which are more redistributive by design, but hardly distinguish between taxes and social contributions.

b) Efficiency

Taxing richer taxpayers more heavily in order to finance transfers to poorer ones may discourage efforts to earn a higher income through participation in the labor force and through longer working hours. In an open economy, heavy taxation may also encourage wealthy households and companies to relocate their tax residence abroad.

In contrast with redistribution, the appropriate variable to assess incentives to work is not the average tax rate, but the *marginal tax rate**, i.e., the fraction of a marginal increase in income that is captured by the tax system. Formally, if $T(R)$ is the tax bill T as a function of before-tax income R, the marginal rate is $T'(R)$ while the average rate is $T(R)/R$. Most personal income tax schedules directly specify marginal tax rates per income bracket.

Strikingly, a tax system can be progressive with a constant marginal rate, if there is a *basic allowance**, i.e., if the first n units of income are exempted from taxation, which is generally the case for the personal income tax. In practice, however, it is difficult to achieve significant redistribution through the tax system with a single marginal tax rate, as will be illustrated in section 7.3.

In most advanced economies, the combination of increasing personal income tax (PIT) rates as a function of income, and of means-tested transfers at the low end of the income distribution, result in *net* marginal tax rates[17] being a U-shaped function of income: Very-low-income households receive social transfers (minimum income, housing, and family benefits) that fall when their income rises, resulting in high marginal net tax rates; the marginal rate also increases at the higher end of the income scale due to PIT progressiveness. The effective marginal tax rate is often higher for low incomes than for higher ones: The discouraging effect of the tax and transfer system is more marked for low-income households, creating *poverty traps**. This feature is illustrated in figure 7.10 in the case of France in 2010. The graph shows the net marginal tax rate of a single-worker couple with two children. The net marginal rate is very high for households earning the minimum wage. It falls dramatically above the minimum wage, before recovering when the household loses housing benefits and starts paying the PIT. It then rises gradually as the income reaches successive income brackets, to arrive at a maximum of 40% of net wage for the highest income bracket.

There is a consensus among economists that high net marginal rates for low incomes have a negative impact on work incentives and lead to

17. The net marginal tax rate is the marginal fall of net income (including the loss of means-tested transfers) following a marginal rise in market income.

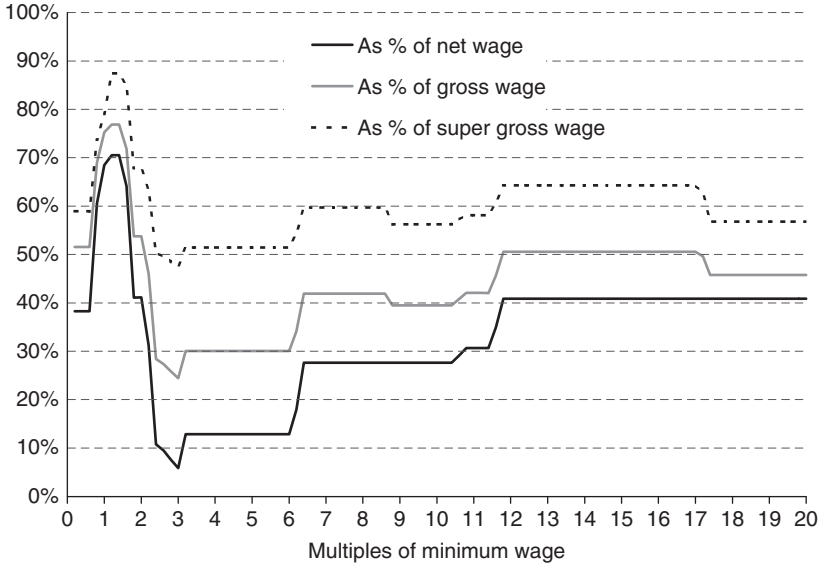


Figure 7.10 Marginal net tax rate for a single-worker couple with two children aged 3-6 in France in 2010.
Source: French Treasury, using "Paris" model.

poverty traps. At the other end of the income scale, the debate on the impact of high marginal tax rates on economic efficiency is more open. The direct impact of taxes on labor supply is probably limited. The risks of discouraging investment in human capital and of encouraging highly qualified workers to move abroad are more significant.

7.2 Theories

Like in other policy areas, the theory of tax policy covers both a positive and a normative dimension. The positive dimension consists primarily in identifying which tax base will finally bear the burden of taxation, measuring the loss of economic efficiency due to distortionary taxes or, conversely, the gain due to targeted taxes (such as environmental taxes). In turn, the normative dimension of tax theory involves laying down guidelines for designing the tax system in an optimal way given social preferences in terms of the efficiency–redistribution trade-off.

7.2.1 Tax incidence on a specific market

A first major insight from tax theory is that taxation is rarely borne by the particular taxpayer that writes the check or orders the bank transfer to the tax administration. For instance, suppose that the labor supply is strictly fixed

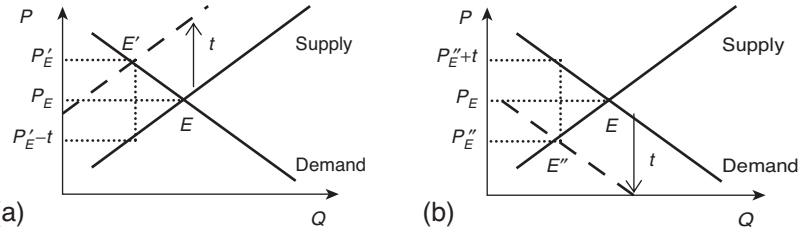


Figure 7.11 A tax on supply or on demand. a) Tax on supply, b) tax on demand.

in a specific place, or for a specific skill: The quantity of hours supplied by workers is constant whatever the wage offered by (competitive) firms. In this particular case, a rise in employers' social contributions will ultimately be borne by employees, since employers will compensate higher tax costs with lower net wages paid to employees. *Ex post*, the fall in net wages has no impact on labor demand since the cost of labor remains unchanged. Although employers pay the contributions, they do not support the tax burden.

More generally, consider a market where supply is positively related to the unit price while demand is negatively related to the price level. This is represented in figure 7.11, where quantities are noted Q and prices P. Market equilibrium is obtained when supply and demand are equal, i.e., where the supply and demand schedules intersect. In the absence of taxes, this corresponds to point E.

Now let us introduce a proportional tax t. The tax can be either a *specific tax** (a fixed amount per volume unit, for instance per ton or gallon), or an *ad-valorem tax** (a fixed percentage of the unit price). Energy taxes are generally of the former type whereas VAT or social contributions are of the latter type. Both types of taxes have similar effects in perfect competition, but react differently to inflation. Here we consider a specific tax, which is easier to tackle graphically (box 7.2 discusses the *ad-valorem* case).

On the left-hand-side of figure 7.11, the tax t is formally levied on supply. In order to compensate for the tax they have to pay, suppliers require a higher price for any level of production: The supply curve moves upwards by t. At the initial before-tax equilibrium price, there is now excess demand, since suppliers are no longer willing to supply the same quantity at this initial price. The market equilibrium moves from E to E', where the quantity is lower and the price paid by consumers is $P'_E > P_E$, whereas the price received by suppliers (after the tax has been paid) is $P'_E - t$. As is apparent in figure 7.11, the tax is partly borne by the demand side since the market price has increased due to the tax. The steeper the demand schedule, the stronger the price increase, hence the greater the share of the tax that is eventually borne by the demand side. In the extreme case where demand is totally rigid (a vertical demand curve), the tax levied on supply is entirely borne by the demand side, since the

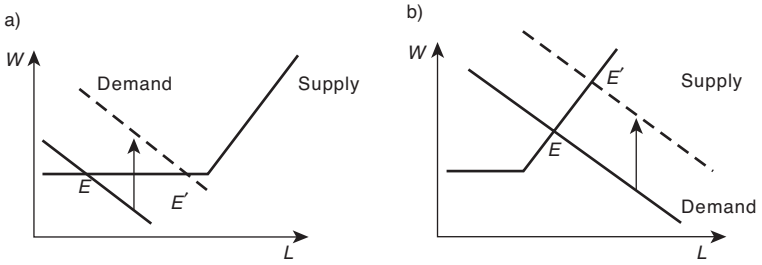


Figure 7.12 Effect of a cut in employers' social contributions depending on the wage level. a) Minimum wage, b) higher wages.

before-tax price received by suppliers remains constant at P_E: *Tax incidence** is on the demand side.

These results have far-reaching practical implications. Employers' contributions, for example, are borne mainly by employees if labor supply is relatively more rigid than labor demand and by employers in the opposite case. In practice, the labor supply schedule is often considered to be bent. At the lower end of the wage scale, there is often a minimum wage (or, equivalently, a back-stop provided by social benefits to the unemployed), and labor supply is perfectly flexible at this wage level. Any increase in employers' contributions will then be borne by employers since they are unable to reduce net wages paid to employees. Symmetrically, a cut in social contributions for low-paid jobs, which has been extensively used in the EU to reduce unemployment, is not passed on to net wages W, which indeed creates an incentive for employers to hire more workers (rise in L, see figure 7.12). For higher wages, labor supply becomes an increasing function of the net wage. For instance, a higher wage may be needed to persuade the partner of an already-working person to take a job. The scarcity of some qualifications can also make labor supply steeper. In that case, a cut in social contributions is clearly shared between employers and employees; moreover, little distinction can be made between employers' and employees' contributions, since both are eventually shared between employers and employees.

Another interesting application of tax incidence relates to consumption taxes (figure 7.11, right-hand-side). The tax shifts the demand schedule downward and the incidence of the tax again depends on the relative slopes of the demand and supply curves. For instance, because of a low price-elasticity of fuel demand, taxes on gas are mostly paid by consumers. Conversely, taxes on manufactured goods are more likely to be shared between consumers and producers because demand for these items is more elastic to the price. Price elasticities are crucial, for instance, when an increase in VAT rates is at stake. If demand is weakly elastic to prices, the VAT increase will be quickly passed on to consumer prices. In the short run, households' purchasing power will fall. In the longer run, wages will likely increase (depending on the slope of the labor-supply curve), in turn feeding inflation. If demand elasticities vary

across products, a rise in VAT can also have a negative redistributive impact: Subsistence items, for which demand is rather inelastic to the price, are likely to suffer a larger price increase than recreation services.

Finally, tax incidence analysis can be applied to tariffs. Protectionist governments impose tariffs on imports as a way of protecting local production and hence, or so they think, the jobs and purchasing power of local workers. But economists view tariffs on imports as the combination of a subsidy to production *and* a tax on consumption. If the price of a product is given internationally, i.e., if the supply curve is horizontal, a t% tax on imports raises by t% the local price for producers and consumers alike: It is fully passed onto consumption prices, which reduces households' purchasing power for all goods, including locally produced goods. Thus, even a unilateral tariff cut is beneficial to households' welfare, a result that some policymakers find hard to believe.

Box 7.2 Tax Incidence: The *Ad-valorem* Case

Most taxes are *ad-valorem* rather than specific, i.e., they are proportional to values rather than to volumes. Suppose that demand Q^d and supply Q^s on a market are linear functions respectively of the price paid P^d and of the price received P^s:

$$Q^d = a_d - b_d P^d, \quad a_d, b_d > 0 \tag{B7.2.1}$$

$$Q^s = b_s P^s, \quad b_s > 0 \tag{B7.2.2}$$

Suppose that an *ad-valorem* tax t is levied on supply: $P^S = P^d(1-t)$. Market equilibrium is given by the equalization of supply and demand:

$$p^d = \frac{a_d}{b_d + b_s(1-t)} \quad \text{and} \quad p^s = \frac{a_d}{b_d/(1-t) + b_s} \tag{B7.2.3}$$

Alternatively, suppose that the tax t is levied on demand: $P^d = P^s(1+t)$. Market equilibrium now yields:

$$p^d = \frac{a_d}{b_s/(1+t) + b_d} \quad \text{and} \quad p^s = \frac{a_d}{b_s + b_d(1+t)} \tag{B7.2.4}$$

For a relatively small tax, the impact on P^d and P^s is the same whether the tax is raised on demand or on supply (since, for a small t, $(1-t) \approx 1/(1+t)$). Tax incidence only depends on the parameters b_s and b_d, which represent the sensitivity of supply and of demand to the price. If supply is not very flexible (small b_s), the price paid by the demand side is little affected by the tax, which therefore weighs on the supply side, whatever the practical organization of tax collection. In contrast, if demand is not very flexible (small b_d), the price received by the supply side is little affected by the tax, which is mostly passed on to the demand side.

More generally, let us denote P the pre-tax price. An *ad-valorem* tax t^d is levied on demand, and an *ad-valorem* tax t^s is levied on supply. The market equilibrium is:

$$Q^d[P(1+t^d)] = Q^s[P(1-t^s)] \tag{B7.2.5}$$

We now study the impact of tax variations on the equilibrium price P. Differentiating (B7.2.5), we get:

$$\frac{\partial Q^d}{\partial P}(1+t^d)dP + P\frac{\partial Q^d}{\partial P}dt^d = \frac{\partial Q^s}{\partial P}(1-t^s)dP - P\frac{\partial Q^s}{\partial P}dt^s \tag{B7.2.6}$$

Denoting ε^s the price elasticity of supply (and ε^d the price elasticity of demand), with $\varepsilon^s, \varepsilon^d > 0$, we get the reaction of the price level to variations in tax rates:

$$\frac{dP}{P} = \frac{\varepsilon^s dt^s - \varepsilon^d dt^d}{(1-t^s)\varepsilon^s + (1+t^d)\varepsilon^d} \tag{B7.2.7}$$

If demand is inelastic to prices ($\varepsilon^d = 0$), a demand-tax increase $dt^d > 0$ has no impact on the pre-tax equilibrium rate P, meaning that the after-tax price rises in due proportion of the tax: The tax increase is entirely borne by the demand side. Conversely, if demand is infinitely elastic ($\varepsilon^d = \infty$), or if supply is inelastic ($\varepsilon^s = 0$), an increase in t^d leads to a proportional fall in the pre-tax price:

$$\frac{dP}{P} = -\frac{dt^d}{1+t^d} \tag{B7.2.8}$$

In this case, the rise in the demand tax is passed on to the supply side. A symmetric reasoning applies in the case of a supply tax.

7.2.2 Social losses and distortions related to taxation

The second policy lesson from tax theory is that, except if it is lump-sum, introducing a tax within a "perfect" market involves a social loss.[18] Relative prices are changed by the tax and they no longer carry the correct information on relative scarcity. For instance, a tax on consumption raises the price paid by the consumer. Accordingly, the latter reduces his or her consumption and his or her utility declines. The fall in consumption triggers a fall in the before-tax market price. Since both the unit price and the quantity sold decline, the producer's profit is reduced. And for the government, there is tax revenue that

18. The concept of social loss is discussed in chapter 1. A social loss appears when there is a fall in social welfare. In this section, social welfare is approximated by the sum of agents' surpluses. Producers' surplus is equal to aggregate profit. Consumers' surplus is the difference between the disposition to pay and the actual market price, for each unit of good. Lastly, the surplus of the public sector is equal to its tax revenue.

can be used to compensate both consumers and producers through lump-sum transfers, but it can be shown (see box 7.3) that the tax revenue does not cover their respective losses, leading to a net social loss.

Box 7.3 Computing the Social Loss

In the previous section, we have seen that taxation introduces a discrepancy between the price paid by consumers and the price received by suppliers. If demand and supply are not rigid, taxation also reduces the quantity produced and exchanged on the market, whether the tax is actually paid by the demand side or by the supply side. Using a simple surplus analysis, figure B7.3.1 measures the resulting social loss. In the absence of a tax, the quantity produced and exchanged is Q_0 and the market price is P_0. In the presence of a tax, output falls to Q_1 and there is now a difference between the price paid by the demand side (here, consumers) P_1^d and that received by the supply side (producers) P_1^s. Table B7.3.1 derives the surplus of consumers, producers, and the government. The tax induces a social loss because the quantity produced and consumed falls. Even if the tax proceeds are redistributed in a lump-sum way (to avoid additional distortions), this is not enough to compensate for the loss incurred by both consumers and producers. The social loss or *deadweight loss** is represented in the figure B7.3.1 by the $C + E$ triangle, called the *Harberger triangle**.[a] The deadweight loss L hence can be measured by the surface of the $C + E$ triangle. It is equal to the base of the triangle (i.e., the tax rate t) multiplied by half the height of the triangle, the latter being:

$$Q_0 - Q_1 = t\left[\frac{\varepsilon^S \varepsilon^D}{\varepsilon^S + \varepsilon^D}\right]\frac{Q_0}{P_0} \qquad \text{(B7.3.1)}$$

where ε^s and ε^d are the price elasticities of supply and demand. Hence, the social loss is:

$$L = \frac{1}{2}t^2\frac{Q_0}{P_0}\left[\frac{\varepsilon^S \varepsilon^D}{\varepsilon^S + \varepsilon^D}\right] \qquad \text{(B7.3.2)}$$

L is higher the higher the price elasticities of supply and demand, and the higher the tax rate. Note that L depends quadratically on t. Assuming a compensated elasticity (see p. 562) of 0.4, Feldstein (2008) evaluates the deadweight loss of a proportional increase of all taxes to be as high as 76% of the incremental revenue.

One practical implication is that, unless taxation aims at correcting specific market distortions (such as pollution externalities), one should avoid elastic tax bases on efficiency grounds. Another practical implication is that a large tax has proportionally more impact on welfare than a small one. This argues for a range of small taxes rather than a single large tax.

However, tax-collection costs generally include a fixed cost, which argues against a proliferation of small taxes.

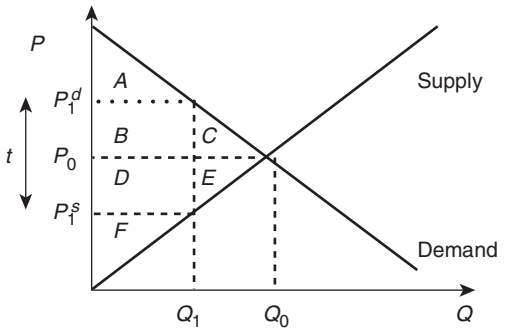


Figure B7.3.1 Taxation and surpluses.

Table B7.3.1
Social loss related to taxation

Surplus	Without tax	With tax	Difference
Consumers	A + B + C	A	− (B + C)
Producers	D + E + F	F	− (D + E)
Government	0	B + D	+ (B + D)
Total	A + B + C + D + E + F	A + B + D + F	− (C + E)

This methodology can be used to assess in monetary terms the deadweight loss of virtually any kind of tax. For instance, Hufbauer and Elliott (1994) have assessed the cost of protection on orange juice in the US. In 1990, the import duty on orange juice was 20% in *ad-valorem* equivalent. They have found the deadweight loss to amount to 70 million dollars, equivalent to 13% of domestic orange juice consumption. This was the sum of a 281 million dollar loss for the US consumer, a 101 million dollar surplus for US producers, a 145 million dollar government revenue, other effects worth 35 million dollars. When moving away from a single good, however, a partial equilibrium analysis may no longer be appropriate, since income effects, with possible spillovers across markets and intertemporal effects, should be accounted for.

[a]For an historical perspective on the Harberger triangle, see Hines (1999a).

a) The Ramsey rule

As shown in box 7.3, the deadweight loss is higher the larger the elasticity of supply and demand to prices, because the tax will have more impact

on quantities. Therefore, minimizing social losses related to taxation requires taxes to be levied on the least elastic goods and services. The Ramsey rule (1927) more precisely states that for the government to minimize deadweight losses while raising a given amount of tax revenue, the tax rate on each market needs to be inversely proportional to the *compensated price elasticities** of supply and demand. A compensated price-elasticity is the variation of supply or demand following a 1% variation in the price level holding income constant (i.e., compensating for the variation of income due to the price variation). This rule can be written as:[19]

$$t = k\left(\frac{1}{\varepsilon^s} + \frac{1}{\varepsilon^s}\right) \tag{7.1}$$

where t is the *ad-valorem* tax rate, k refers to the total amount of taxes to be collected, ε^d the compensated price-elasticity of demand and ε^s the compensated price-elasticity of supply. Note that the Ramsey rule aims at levying taxes while introducing as few distortions as possible. Hence, it concentrates on the allocation function, neglecting the redistribution function of tax policy. Indeed, applying the Ramsey rule can lead to unfair policies. For instance, it involves taxing more the least reactive tax bases, e.g., unskilled labor rather than capital or skilled labor, or bread rather than perfumes and health expenditures rather than theater tickets.

b) "Too much taxes kill taxes": The Laffer curve

Another consequence of the surplus analysis developed in box 7.3 is that the tax revenue is not a monotonous function of the tax rate: A tax-rate increase has two opposite effects on tax receipts. On the one hand, each unit of the tax base is taxed more heavily, which raises revenues. On the other hand, the tax base is reduced by the tax increase, which at a given tax rate cuts revenues. The net effect again depends on the price-elasticities of supply and demand (see box 7.4). Starting from no taxation at all, a tax increase raises tax revenues, but less and less so as the tax rate increases. After a certain threshold called the *revenue-maximizing tax rate**, a rise in the tax rate *reduces* tax receipts, because the positive impact of the tax increase is over-compensated by the reduction in the tax base. The revenue-maximizing tax rate can be very high when demand is inelastic to prices, but it can be low for very elastic tax bases, e.g., internationally mobile tax bases (see section 7.2.6).

Box 7.4 Tax Rates and Revenues

Suppose an *ad-valorem* tax t is levied on the supply side. The tax revenue $R(t)$ is the product of output Q by the difference between the price paid

19. See Salanié (2003).

by the demand side P^d and the after-tax price received by the supply side P^s:

$$R(t) = Q(P^d - P^s) \tag{B7.4.1}$$

With the simple supply-and-demand functions of box 7.2, one obtains:

$$R(t) = \left(\frac{a_d b_s(1-t)}{b_d + b_s(1-t)}\right)\left(\frac{a_d t}{b_d + b_s(1-t)}\right) \tag{B7.4.2}$$

It can be seen that $R(0) = 0$ and $R(1) = 0$. Between these two extreme values of t, there is a single rate t^* that maximizes the tax revenue. This rate is:

$$t^* = \frac{b_d + b_s}{2b_d + b_s} \tag{B7.4.3}$$

Beyond t^*, the tax revenue decreases when t increases. It can be noted that t^* is higher if a limited proportion of the tax is passed on prices (high b^s), and if the increase in the price does not discourage too much demand (small b^d).

This inverted-U-shaped curve was popularized in the 1970s by Arthur Laffer after he had supposedly sketched it on a napkin at a December 1974 working lunch (Wanniski, 2005). The idea was not new (it had already been hinted at by David Hume and by John Maynard Keynes) but Laffer surprised his contemporaries by declaring that, in view of the high tax pressure in the US at that time, a cut in the income tax rate was likely to *raise* tax receipts (Laffer, 2004). Put differently, he was supposing that the income tax rate was lying on the downward-sloping section of the inverted U-shaped *Laffer curve** (cf. figure 7.13). His argument had a strong influence on President Ronald Reagan's and President George W. Bush's tax-cutting policies during the first half of the 1980s and in the early 2000s, respectively. In all cases, however, tax cuts led to steady increases in the budget deficit, showing that the economy was not in the downward-sloping section, but rather on the upward-sloping section of the Laffer curve.

A more compelling illustration of the Laffer curve is the Russian personal income tax reform of 2001, which involved a sharp fall in the top marginal tax rate of the personal income tax, from 30% to 13%. Related tax receipts eventually rose by 25% in real terms (figure 7.14). Even in this case, however, it is not clear whether tax revenues increased due to the positive impact of reduced rates on the supply side, or due to better tax compliance encouraged by reduced rates but also to accompanying enforcement measures (see Ivanova et al. 2005).[20]

20. There is also some evidence of a Laffer curve for corporate income tax, especially since multinational firms may shift profit abroad as the tax rate increases relative to the foreign one. See Clausing (2007) and Devereux (2006).

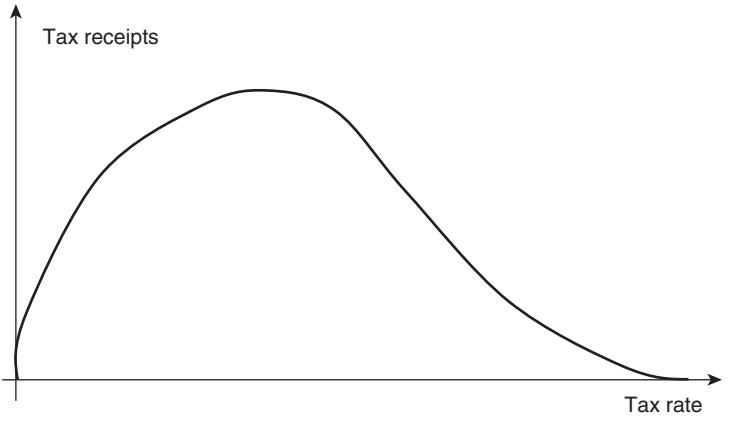


Figure 7.13 The Laffer curve.

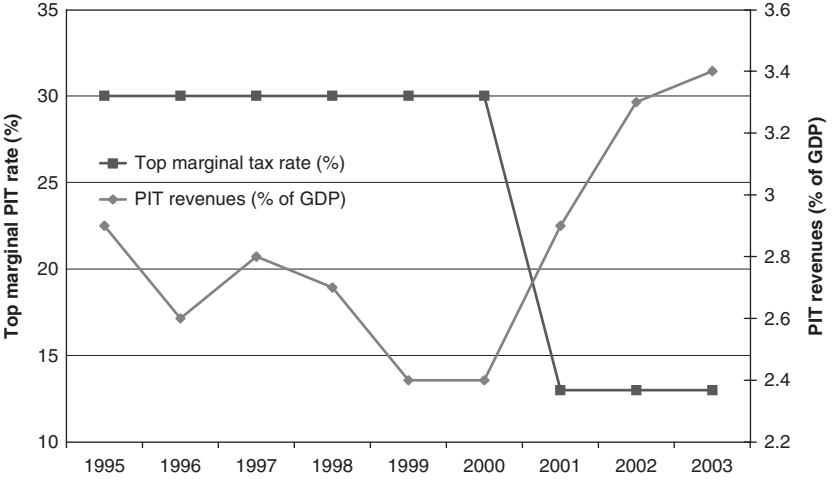


Figure 7.14 A Lafferian moment: The 2001 Russian personal income tax reform. Source: Ivanova et al. (2005).

Note: PIT stands for personal income tax.

Whatever its empirical relevance, the Laffer curve acts as a backstop for decision-makers, threatening them with reduced tax revenues if they raise rates beyond a certain threshold. However, the Laffer curve does not constitute any fiscal "theory," which would need detailed modeling of microeconomic behaviors in each area of taxation. Neither does it provide any operational guide: In the absence of a precise specification of behaviors, one cannot determine whether the average tax rate of the economy is higher or lower

than its "optimum" level, and therefore whether a tax rise would lead to higher or lower tax revenues.

7.2.3 Tax incidence in general equilibrium

Up to now, we have focused on a single specific market and used a partial equilibrium approach. This is obviously a limitation since the behavior of, say, consumers, depends on the variation of all relative prices (including consumption prices, but also wages, interest rates, etc.), as well as on the variation of their income. Hence, taxation on one specific market alters behaviors on other markets by modifying relative prices (substitution effects) and purchasing power (income effects).

a) A simple example

Figure 7.15 illustrates the need for general-equilibrium reasoning on the specific case of a consumer allocating his nominal income R between two goods in quantities C_1 and C_2, respectively, providing a utility $U(C_1, C_2)$. On the graph, the choice of the consumer is represented by point E where the budgetary constraint and the family of iso-utility curves (indifference curves) are tangent. *Ex ante*, i.e., before price adjustments occur, the introduction of a tax on good 1 moves the budgetary constraint downward and clockwise (around the point corresponding to zero consumption of good 1). Utility maximization then leads the consumer to replace good 1 with good 2 in his or her consumption basket, because the relative price of good 1 increases. However, the purchasing power of his or her income R is reduced by the tax, which leads to a fall in the consumption of both goods. The net effect on the consumption of good 2 is ambiguous. Suppose that demand for both goods decreases. Their after-tax prices will fall, which eventually will shift the budgetary constraint to the North-East of the graph. By ignoring this effect, one would over-estimate the impact of the tax on the market for good 1.

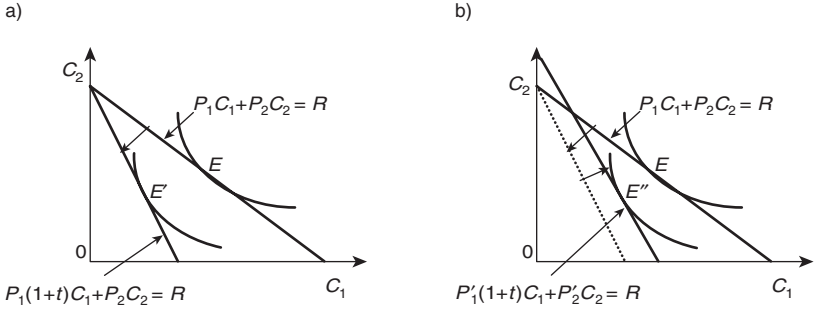


Figure 7.15 Effect of a consumption tax in a two-good model (tax on good 1). a) Before price adjustment, b) after price adjustment.

b) An application to taxes on labor and capital income

Taxing capital income amounts to raising the price of deferred—relative to immediate—consumption. If the substitution effect dominates (households prefer consuming immediately because it is less expensive than consuming tomorrow), the saving rate falls; conversely, if the income effect dominates (households need to save more today in order to maintain a given level of consumption tomorrow), the saving rate increases. The model detailed in box 7.5 shows that to avoid distorting the trade-off between consumption and saving, it is more appropriate to tax consumption rather than income. The reason is that, when there is a tax on income, saved income is actually taxed twice (first as labor income, and then as capital income). This is why the personal income tax system often involves a lower tax rate on capital income than on labor income.[21] Empirically, the compensated elasticity of the saving rate to the personal income tax appears slightly negative (ranging from 0 to −0.3; see for example Sandmo, 1985). Hence, savings tend to fall slightly through a substitution effect following a saving tax increase. In a closed economy, however, this fall in savings is likely to engineer a rise in its before-tax return. The upward adjustment of the real interest rate then wipes out the impact of the tax increase for savers, but the tax is passed on to companies, which suffer from the rising cost of investment.

Box 7.5 The Impact of Taxes on Savings Income

Consider an individual living two periods. In period 1, he or she is young: He or she works and receives a wage w that is used to pay social-security contributions and a personal income tax at rate t_w, to consume a quantity c_1 of a representative good that we take as the numeraire (which means that its price is equal to 1), to pay a consumption tax t_c and to save an amount s. In period 2, he or she is old, no longer works but consumes the product of his or her savings after paying a tax t_s on savings income and a consumption tax t_c. For simplicity, we assume that there is no bequest. The real interest rate is denoted r. The budget constraint for each period is:

$$\text{Period 1: } (1 + t_c)c_1 = (1 - t_w)w - s \qquad \text{(B7.5.1)}$$

$$\text{Period 2: } (1 + t_c)c_2 = (1 + r(1 - t_s))s \qquad \text{(B7.5.2)}$$

Assume that the individual maximizes an intertemporal-type CES utility function:[a]

$$\underset{s}{Max}\ U(c_1, c_2) = \left(c_1^{\frac{\sigma-1}{\sigma}} + \beta c_2^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} \qquad \text{(B7.5.3)}$$

21. Other justifications include the mobility of capital tax bases, or a double taxation of dividends (which are already taxed at the level of the firm).

where σ is the intertemporal elasticity of substitution ($\sigma > 0$) and β the preference rate for future consumption ($0 < \beta < 1$). The resolution of the optimization program leads to the following level of saving:

$$s = \frac{(1 - t_w)w}{1 + \beta^{-\sigma}(1 + r(1 - t_s))^{1-\sigma}} \qquad \text{(B7.5.4)}$$

The level of saving depends positively on disposable income in period 1, $(1 - t_w)w$. Note that the income tax t_w has a proportional impact on savings by lowering period 1 disposable income. The impact of the after-tax return on saving $r(1 - t_s)$, however, depends on whether the intertemporal elasticity of substitution σ is higher or lower than unity. If $\sigma > 1$, then a rise in the savings tax, by reducing the after-tax return, reduces the level of saving. If $\sigma < 1$, we get the opposite effect as the individual will save more for his/her old age. Lastly, the consumption tax t_c plays no role insofar as it affects consumption in the two periods in the same way.

[a] CES stands for Constant Elasticity of Substitution: The elasticity of substitution does not depend on the amounts consumed at the two periods. See also chapter 6.

On the capital market, households supply their savings (capital supply) while firms look for investment finance (capital demand). Market equilibrium is achieved through real interest-rate adjustment. Capital supply can be thought of as relatively flexible due to a high intertemporal elasticity of substitution, or to the existence of alternative investment opportunities (in foreign markets, or simply in government bonds). In contrast, capital demand is relatively rigid because productive investment depends more on market prospects than on the real interest rate. Consequently, both taxes on savers and on investors are likely to be borne by the demand side, i.e., by firms: The before-tax return on capital has to increase to ensure unchanged after-tax return for savers.

To raise the before-tax return on capital, a firm must reduce its capital stock (assuming, typically, that marginal productivity is a decreasing function of the capital stock). Because the capital stock per worker falls, labor productivity also falls, which leads to a reduction either in wages or (if and when wages hit the minimum-wage floor) in employment (cf. box 7.6). On the whole, the incidence of a tax on savings depends on the relative elasticities of capital supply and of labor supply. The more rigid is the labor supply relative to capital supply, the higher is the share of the tax eventually borne by workers (through lower wages) rather than capital owners (through reduced after-tax return). Empirically, Arulampalam et al. (2007) show that at least 54% of an additional corporate tax is passed through to lower wages and that this proportion even exceeds 100% in the long run.[22]

22. This striking result is based on data for 23000 companies located in 10 countries over the period 1993–2003, i.e., during a period of high capital mobility. In a closed economy, Auerbach

This discussion qualifies the traditional debates on the burden-sharing of taxes between labor and capital: Although a tax on capital raises the price of capital relative to labor, which can involve favorable substitution effects for employment, its primary long-term effect is a loss in labor income. Thus, a tax on savings can ultimately have the opposite impact from what is generally believed (for an argument along these lines, see Feldstein, 2005).

Box 7.6 The General Equilibrium Effect of Capital Taxes

Here we build on the savings model presented in Box 7.5. Suppose that period 1 savings are used to acquire productive capital that the old generation sells to the young one. For simplicity, we assume that there is neither capital depreciation nor demographic growth. As there are only two generations, the young generation must buy the entire capital stock of the economy. The supply of capital by each young person is (see box 7.5):

$$k^s = s = \frac{(1 - t_w)w}{1 + \beta^{-\sigma}(1 + r(1 - t_s))^{1-\sigma}} \tag{B7.6.1}$$

Here we assume that $\sigma > 1$, so that capital supply is an increasing function of the after-tax return. We assume a Cobb–Douglas production function,

$$Y = K^{\alpha} L^{1-\alpha} \tag{B7.6.2}$$

where K represents the capital stock, L employment and $0 < \alpha < 1$. Let us call $y = Y/L$ the output per worker and $k = K/L$ the capital stock per worker. The per-capita level of output and income is $y = k^{\alpha}$. The marginal productivity of capital is αk^{a-1} and that of labor is $(1 - \alpha)k^{\alpha}$. Profit maximization involves equalizing each of these marginal productivities to the corresponding factor cost. If t_{ssc} designates the employers' social contribution rate and t_{cit} the corporate income tax (CIT), and if capital depreciation is ignored, profit maximization leads to:

$$\text{Capital: } (1 - t_{cit})\alpha k^{\alpha-1} = r \tag{B7.6.3}$$

$$\text{Labor: } (1 - \alpha)k^{\alpha} = (1 + t_{ssc})w \tag{B7.6.4}$$

From equation (B7.6.3), it is possible to recover capital demand as a decreasing function of the real interest rate and of the corporate income

(2005) argues that capital owners may bear a large part of a corporate tax increase in the short run because the price of their shares is immediately reduced. In the longer run, the tax is borne by both corporate-capital and non-corporate capital owners, as shown by Harberger (1962). However, the short-run matters, especially because it affects inter-generation redistribution.

tax rate:

$$k^d = \left(\frac{r}{\alpha(1 - t_{cit})} \right)^{1/(\alpha-1)} \tag{B7.6.5}$$

Then, equation (B7.6.4) shows how the wage that firms are prepared to pay depends positively on capital per worker. Together with (B7.6.5), this leads to the following negative relation between the CIT and the wage that firms are prepared to pay, for a given interest rate:

$$w = \frac{1 - \alpha}{1 + t_{ssc}} \left(\frac{r}{\alpha(1 - t_{cit})} \right)^{\alpha/(\alpha-1)} \tag{B7.6.6}$$

The negative impact of the CIT on the wage rate can be moderated by its negative impact on the interest rate (in a closed or a large economy), which triggers substitution from capital to labor, and the final impact depends on the reactions of supply and demand on both the capital and the labor market.

Figure B7.6.1 represents the supply and demand for capital as functions of the real interest rate. A rise in the personal income tax rate (or in employees' social insurance contributions), or a rise of the savings tax, shifts the capital supply curve to the left. The interest rate rises to restore the balance between supply and demand. Conversely, a rise in the corporate tax rate shifts the capital-demand curve to the left: The real interest rate decreases to restore equilibrium. In both cases, capital per worker falls. The result is a reduction in the marginal productivity of labor. Symmetrically, a rise in employers' social insurance contributions (t_{ssc}) reduces employment, and therefore both the productivity of capital and its return. The relative impact of taxation on wages and on capital returns depends on the relative slopes of the supply and demand curves in both markets.

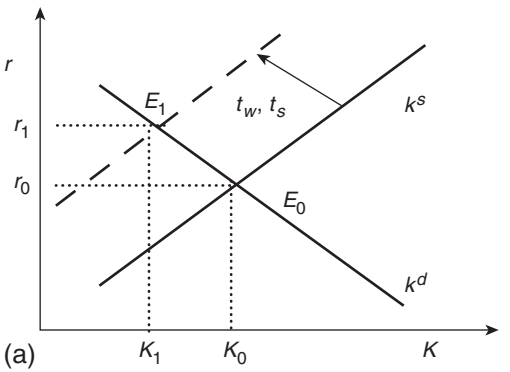


Figure B7.6.1 Impact of various taxes on the market for capital. a) Taxation of households' income.

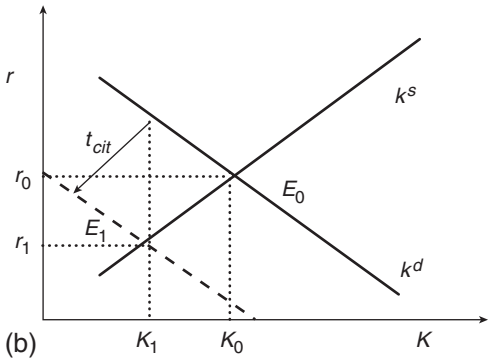


Figure B7.6.1 (Cont'd) b) corporate taxation. Reading: a) A rise in the personal income tax t_w or in the savings tax t_s reduces capital supply k_s, which leads to a rise in the real interest rate r, a fall in the per capita stock of capital k and, ultimately, a fall in labor income; b) a rise in the corporate income tax t_{cit} lowers capital demand k_d, which leads to a decline in the real interest rate but also lowers the capital stock and therefore labor productivity.

7.2.4 Effectiveness versus equity: Optimum taxation

As detailed in the previous subsection, the search for efficiency calls for raising public revenues through a low-rate tax on a broad, relatively inelastic base, and for avoiding multiple taxation of the same base. Moreover, taxes are generally borne by the least flexible tax base, possibly on another market than the one where taxation takes place. Accordingly, it will be difficult to have capital income really contribute to the public budget, which raises difficult questions of tax equity. The literature on *optimum taxation** tries to identify the tax structure that allows the best trade-off between efficiency and equity.[23]

A first approach consists in minimizing the efficiency cost of taxation for a given amount of public revenue. This leads to the Ramsey rule, which recommends taxing the various tax bases in opposite proportion to the compensated elasticities of supply and of demand (cf. supra). An extension consists in considering leisure as a "good" whose price is its opportunity cost (namely the consumption lost by an individual who spends an hour of leisure instead of working). The Ramsey rule would imply (Feldstein, 1978) that items that are close substitutes for leisure should be lightly taxed lest

23. For a review, see Slemrod (1990). See also Stern (1987) and Auerbach and Hines (2002).

individuals choose to substitute leisure for work. This might lead to taxing savings (i.e., deferred consumption) rather than current consumption, which can be thought a closer substitute for leisure. This principle also justifies a light taxation of personal services such as gardening, child care, or housekeeping, because these activities are close substitutes for leisure: Assuming that working time is flexible, taxing these activities too heavily might lead households to reduce their working time in their usual professions in order to stay at home more, and to carry out these tasks themselves.

However, as already mentioned, the Ramsey rule can lead to an inequitable distribution of the fiscal burden. A more elaborate approach, which has inspired most of the literature on optimum taxation, consists in introducing an equity objective alongside the efficiency one. For example, the optimum tax rate on luxury goods will be higher than implied by the Ramsey rule, whereas the optimum tax on necessity goods will be lower.

Through James Mirrlees's pioneering work (1971), the optimum taxation literature first addressed personal income taxation, which is usually considered as one of the main policy instruments for interpersonal income redistribution (even though, as discussed earlier, income redistribution also increasingly takes place through nontax instruments). Mirrlees defines the social utility function as a weighted sum of individual utilities, with weights inversely proportional to individual incomes, which amounts to giving more consideration to the poorest. This social utility function is maximized under two constraints: A public *income constraint* (i.e., the tax revenue to be collected), and an *incentive constraint* that recognizes the impact of taxation on the incentive to work.

Assuming that each individual's income is equal to his/her marginal productivity, redistribution requires taxing more heavily higher productivity individuals; but this is likely to discourage the most productive individuals from working, and therefore to reduce tax revenues by shrinking the tax base. An optimum income tax profile is derived from this trade-off between equity and efficiency (cf. box 7.7).

Box 7.7 Optimum Taxation[a]

Mirrlees' general result (1971) can be summarized in the following way: The optimum marginal tax rate for an individual with productivity w is $T'(w)$ such that:

$$\frac{T'(w)}{1 - T'(w)} = E(w)R(w)H(w) \qquad \text{(B7.7.1)}$$

where $E(w)$ is a decreasing function of the elasticity of labor supply, consistent with the Ramsey principle. $R(w)$ is the weight allocated by the government to individuals with productivity w in the social utility function ($R'(w) < 0$). Lastly, $H(w)$ reflects the income structure. It is

a decreasing function of the number of individuals with productivity w, and an increasing function of the number of individuals with productivity higher than w:

$$H(w) = \frac{1 - F(w)}{wf(w)} \qquad \text{(B7.7.2)}$$

where f is the statistical distribution of w and F its cumulated distribution: The marginal rate of taxation of individuals with productivity w must not be too high if there are many such individuals (a high $f(w)$), in order to dampen what could be a massive discouraging effect on labor supply; conversely, a rise in the marginal tax rate of individuals with productivity w is appropriate if there are many individuals with a productivity higher than w ($1 - F(w)$ high), because these individuals then contribute significantly to the budget without facing a disincentive to work.

The first results by Mirrlees (1971) point to a roughly linear optimum tax rate with relatively low marginal rates (between 20% and 30%). However, these results strongly depend on assumptions regarding the social utility function or the elasticity of labor supply. It is not easy to reconcile them with the typical U-shaped profile of marginal tax rates observed in OECD countries (see above, figure 7.10). Refined versions of the Mirrlees model have introduced a lower substitutability between consumption and leisure at the lower end of the income scale, which justifies higher marginal rates for low-income individuals, and have thus reached results that are more consistent with observed facts. Accounting for the clustering of labor productivities in the middle of the income scale also helps in recovering a U-shape curve, since it is optimal to reduce taxation on the most numerous groups of individuals in order to limit the tax-induced fall in labor supply.

[a] See also Salanié (1998).

On the whole, optimum taxation theory provides a better understanding of the efficiency–equity trade-off but hardly provides operational guidance to governments contemplating a tax reform. As underlined by Slemrod (1990), measuring the elasticity of labor supply or the degree of substitution between consumption and leisure is particularly difficult; and optimum taxation models rarely argue for strongly progressive tax profiles, except when they assume very low elasticities of substitution between consumption and leisure. Furthermore, these models are highly stylized. For instance, they do not distinguish between the individual and the household, even though the composition of a household is crucial for the elasticity of the individual labor supply and for the individual's utility. Lastly, optimum taxation theory neglects the costs of tax collection, which include the management of the tax

system but also the prevention of fraud and of tax avoidance:[24] Tax collection costs depend on the instruments used and on the nature of taxation, and therefore need to be taken into account in the design of an optimum taxation system (Slemrod and Yitzhaki, 2000).

7.2.5 Corrective taxation

So far, taxes have been found to be detrimental to economic efficiency because they distort choices by changing relative prices between goods, between consumption and leisure, or between present and future consumption. However, this approach is only valid in a flawless economy with no market failures such as imperfect competition, externalities, asymmetrical information, etc. In the presence of market failures, changing relative prices can actually bring the economy *closer* to efficiency. Taxation can in such cases substitute for other policies such as regulations or codes of conduct.

This idea goes back to the 1920s when Arthur Pigou (1920) proposed the introduction of a tax on London chimney emissions in order to fight the infamous "smog." This involved bridging the gap between the *private cost** of emissions, incurred by the agents who were responsible for them, and their *social cost**, which includes the damage caused to other agents (*polluter-payer principle**).[25] This principle has now been introduced in many countries. In the same way, a congestion charge was introduced in London in 2003 for motorists choosing to enter the city center. This £5 (later £8) fee seems to have been successful in reducing congestion.[26] In that particular case, the externality that was thus "internalized" through taxation was not pollution, but rather congestion and its associated damages (noise, commuting delays . . .). The effectiveness of so-called *Pigovian taxes** hinges on equalizing the marginal cost of emission reduction across polluters (see box 7.8): All polluters will reduce their emissions up to the point where the marginal cost of further reductions is equal to the tax (beyond that point, they will prefer paying the tax rather than incurring the costs of further reducing their emissions). This means that the firm that faces the lowest marginal cost of cutting emissions will reduce its emissions more than a firm facing a higher marginal cost. This is efficient for the economy as a whole, in contrast with the imposition of a uniform emission limit for any firm.

However, Pigovian taxes affect emissions only indirectly, through changes in marginal costs. Their success relies on adequately assessing the social cost of damage and the reactivity of behaviors to price variations. If these

24. See box 7.12.

25. The polluter-payer principle is in fact very old. From the Middle Ages to the French Revolution, a so-called *pulverage charge** existed in Dauphiné and in Provence (France). This tax was charged by villages crossed by transhumant herds to compensate for the dust raised by their passage.

26. Congestion was reduced by 26%, according to Transport of London's fourth annual report on the congestion charge, June 2006 available on: http://www.tfl.gov.uk.

Box 7.8 The principle of a Pigovian tax

The idea behind the Pigovian tax is to have polluters internalize the cost of pollution. Assume for instance that households consume a good in quantity y, the production of which releases an amount e of pollution that deteriorates households' welfare. Denoting by $U(y, e)$ the utility function, we have:

$$\frac{\partial U(y, e)}{\partial y} > 0; \frac{\partial U(y, e)}{\partial e} < 0 \qquad \text{(B7.8.1)}$$

The production cost, $C(y, e)$ is an increasing function of the quantity produced and a decreasing function of the pollution released; The marginal cost is increasing in y but decreasing in e. Assuming that the amount of pollution is limited to $\bar{e}$, we have:

$$\frac{\partial C(y, e)}{\partial y} > 0; \frac{\partial C(y, e)}{\partial e} < 0; \frac{\partial^2 C(y, e)}{\partial y^2} > 0; \frac{\partial^2 C(y, e)}{\partial y \partial e} < 0; e \leq \bar{e} \qquad \text{(B7.8.2)}$$

In the decentralized, perfect competition equilibrium, polluting emissions are at their upper bound and the firm charges a price p equal to its marginal cost with this maximum pollution level:

$$\underset{y,e}{Max}\ \pi = py - C(y, e) \Rightarrow p = \frac{\partial C(y, \bar{e})}{\partial y} \qquad \text{(B7.8.3)}$$

The firm releases as much emissions as possible because it does not internalize the social cost of pollution. A Pigovian tax can however be levied on each unit of production to make the firm internalize the impact of pollution on households' utility. The rate of the tax t must be equal to the marginal disutility of pollution:

$$t = -\frac{\partial U(y, e)}{\partial e} \qquad \text{(B7.8.4)}$$

With a Pigovian tax in place, the private cost of production becomes equal to its social cost.

parameters are known, then it is possible, using a tax, to exactly reach the desired quantitative objective (for example, a given reduction of pollution). However, if these parameters are uncertain, the quantitative results from a tax will also be uncertain (cf. figure 7.16).

Another solution for correcting externalities is to rely on regulations. In the London case, for instance, it could have been decided that only cars with even plate numbers would be allowed to enter inner London on even days and odd-plated cars on odd days.[27] The quantity of vehicles entering the city

27. Other schemes are, of course, possible for regulating traffic in cities, such as granting access only to emergency vehicles, buses, and delivery vehicles during certain hours.

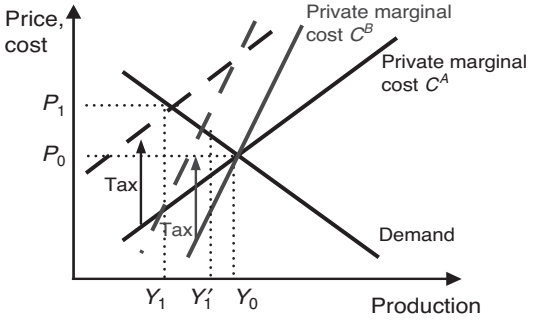


Figure 7.16 Impact of a Pigovian tax.
Reading: The tax raises the private marginal cost curve C^A upward and to the left, leading to a production decline from Y_0 to Y_1. An error on the marginal cost slope (e.g., C^B instead of C^A) limits the production fall to Y_1'. The reduction in polluting emissions is also lower, although the tax cost per unit is the same.

would then have been certain (in our case, a 50% drop in traffic could have been expected), but the cost of the regulation would have been uncertain. As mentioned above, this would not have been a cost-efficient way of reducing the traffic.

Another alternative to taxation is based on the recognition that externalities emerge as a result of a missing market (for instance, a market for clean air or for smooth-flowing traffic) and consists in introducing this new market. For example, issuing *tradable emission permits** or tradable traffic permits would allow both control of the total quantity of pollution (or of traffic), related to the volume of available permits, and limiting the cost of pollution (or traffic) reduction, by concentrating reduction efforts on those who will suffer less. For instance, the transport authority of a big city could allocate a given volume of inner-city traffic permits to residents for one semester corresponding to, say, a 25% reduction in traffic compared to past figures. Those who use public transportation would have the opportunity to sell their permits on a market, and those who did not have enough permits (or did not live in this area) could buy these permits at market price. One advantage of this system is that the transport authority would not need preliminary knowledge on the relationship between price and traffic; a second advantage is that those individuals who are able to use public transport would gain from selling their permits. Such tradable permit schemes with an initial endowment are called *cap-and-trade** systems. Prominent examples are the US sulfur dioxide trading system established under the 1990 Clean Air Act to combat acid rain, and the EU Greenhouse *Gas Emission Trading System* (*ETS*)* launched in 2005 for carbon dioxide emissions. A cap-and-trade system sets

an overall pollution quantity and lets the price be set at a decentralized level (while a Pigovian tax sets an overall price and lets the pollution level vary). This allows for both controlling the total volume of pollution and directing the effort to those polluters which face the lowest cost of reducing their emissions.

In theory, any problem of externalities can be resolved through negotiation. For example, smokers can be forbidden from smoking, or they can be required to negotiate with nonsmokers the right to smoke in exchange for some compensation. In 1937, in his book *The Nature of the Firm*, Ronald Coase, a British economist and 1991 Nobel Prize winner, stated what has subsequently been called the *Coase Theorem**: As long as all parties are free to bargain, negotiation will deliver an efficient outcome irrespective of legal entitlements. If the law forbids smoking, it is up to the smokers to buy the nonsmokers' indulgence, whereas if smoking is allowed, it is up to the victims to buy pure air from smokers. The Coase theorem is valid only in the absence of (or with limited) transaction costs, e.g. it can fix neighborhood disputes, but cannot solve the global warming problem. Furthermore, the outcome of the negotiation generally depends on the initial allocation of property rights. In the case of industrial pollution, the difficulty is precisely to establish property rights: Do firms have the right to pollute when they produce, in which case the firm has to be subsidized to reduce its emissions, or is the planet entitled to preserve its climate, in which case firms have to pay for the pollution they generate? The Pigovian tax adopts the latter viewpoint whereas the market for tradable emissions permits is more flexible, depending on the initial allocation of permits.

What should be done with tax revenues? There are three possibilities. The first consists in compensating polluters through a lump-sum (or otherwise) transfer, in order not to penalize them unduly or, from a more political-economy perspective, to make the tax more acceptable to them. For example, the Swedish power stations are taxed proportionally to their nitrogen dioxide emissions, but they receive a transfer proportional to their electricity production. This taxation-cum-redistribution scheme allows behavior to be directed toward a reduction of emissions without modifying the net tax burden for the sector as a whole. The second possibility is to use tax revenues to produce public goods, in particular to finance environmental expenditures. This is the option chosen in London, where congestion charge receipts are invested in the city's transportation infrastructure. Finally, the tax revenue can be used to cut other taxes considered as distorting, especially taxes on labor: This allows a *double dividend** to be reaped, since social welfare rises both because of the tax itself (which corrects an externality) and because of the cut in tax-induced distortions. Germany and The Netherlands thus substituted eco-taxes for social insurance contributions (see section 7.3). The very existence of a double dividend is however debated, since (i) under perfect competition on the goods market, the incidence of green taxes is likely to fall on labor, and (ii) the success of a Pigovian tax means that

the tax base will shrink, which makes cuts in social security contributions unsustainable.

7.2.6 Taxation in open economies

It has been argued above that taxes are ultimately borne by the least flexible bases. In an open economy, capital is generally more mobile than goods and especially labor. Hence, the burden of taxation tends to fall on labor and consumption, two relatively immobile tax bases.

a) Capital mobility

When capital is mobile, asset-holders can choose where to invest from among various places. For a given level of risk, they will invest where the return to capital is highest. Perfect arbitrage will lead to a single risk-adjusted capital return worldwide.

Suppose now that a tax is levied on capital earnings. This tax can be raised according to the *resident principle** or according to the *source principle**. In the former case, a household, for example in Germany, earning capital income from the US pays a tax in Germany, either as a component of its personal income tax or separately (through a withholding tax, a tax on capital gains or else an inheritance tax). The after-tax capital return will be uniformly reduced across assets. The household can escape the tax by moving its residence to a low-tax country, or by locating its capital income in a bank account in a country that will not inform the tax administration of its country of residence, thanks to bank secrecy.

In the source-principle case, capital income is taxed in the country where it is earned. For instance, if a German resident holds stock from a US company, his/her capital income will result from the distribution of dividends based on after-tax company profits (in practice, after the corporate income tax, CIT, has been levied). If the tax is lower in the US than in Germany, capital flows to the US, which reduces the before-tax return to capital in the US until the after-tax return is equalized with the rest of the world. *Ex post*, the German asset-holder receives the same rate of capital return from US and German investments (but a different before-tax return).

Whether capital is taxed under the resident principle or under the source principle, arbitrage by asset-holders tends to equalize after-tax returns across countries. This translates into a perfectly flexible capital supply, so that any tax variation falls on the demand side of the capital market, and in turn is passed on to labor (see section 7.2.3).

Under these conditions, a growth-enhancing strategy consists in reducing capital tax rates, in order not only to increase the tax base (through capital inflows), but also, in the corporate tax case, to raise the productive capital stock and the level of employment. Such a strategy has been successfully adopted by Ireland since the 1980s. This is obviously not cooperative: If all countries

cut their corporate tax rates similarly, they will lose tax revenues without attracting foreign capital inflows; although domestic firms in each country will be willing to invest more, the worldwide surge in investment demand will lead to a rise in the interest rate.

This is called *tax competition**. Tiebout wrote the seminal paper on the question in 1956. He concluded that each individual would relocate to the fiscal jurisdiction offering the combination of taxes and government services that would be closest to his/her preferred basket: Individuals would "vote with their feet," and this process would discard inefficient jurisdictions (those offering too few government services for the prevailing level of taxation). In this framework, competition eliminates ineffective jurisdictions, while allowing a diverse level of public good provision fitting the diversity preferences.

This "happy" tax competition scenario, however, may well be overly optimistic. Zodrow and Mieszkowski (1986) have shown that, if all jurisdictions have similar preferences, the provision of government services under such tax competition will be lower than its optimum level in autarky. Indeed, each jurisdiction considers that a rise in its tax rate carries the risk of pushing taxpayers out and does not realize that the other jurisdictions are in the same situation. Thus, no jurisdiction dares fix the tax rate which, adopted by all jurisdictions, would be the optimum (see box 7.9), and all of them overestimate the cost of government services (in terms of lower consumption of private goods).

Assuming now that government services are financed both by a tax on mobile capital and by a tax on immobile labor, it can also be shown that the (insufficient) provision of government services will be mainly (if not exclusively) financed through labor taxation (Bucovestky and Wilson, 1991). Figure 7.17 contrasts the downward trend of corporate tax rates in the EU since 1995 with the upward trend of EU-15 standard VAT rates, which are raised on almost immobile consumption. During and after the 2007–09 global economic crisis, raising VAT rates was often viewed as an efficient way to restore public finances. New EU member states, that suffered sudden stops in foreign capital inflows during the crisis, were the first to hike their VAT rates; Greece, Spain, and other countries followed after a few months.[28]

These traditional results of the literature on tax competition, which predict a "race to the bottom" in corporate tax rates, have been questioned since the late 1990s by the "new economic geography" literature (see Baldwin et al., 2003, and chapter 6 of this book). According to this new research avenue, large, geographically connected countries benefit from agglomeration effects allowing them to maintain higher tax rates without suffering from a relocation of their activities. These agglomeration effects are related to economies of scale, which create an incentive for firms to concentrate their activities in a small number of places, provided that transport costs between production

28. The UK temporarily cut its standard VAT rate in 2009, as part of its fiscal stimulation package.

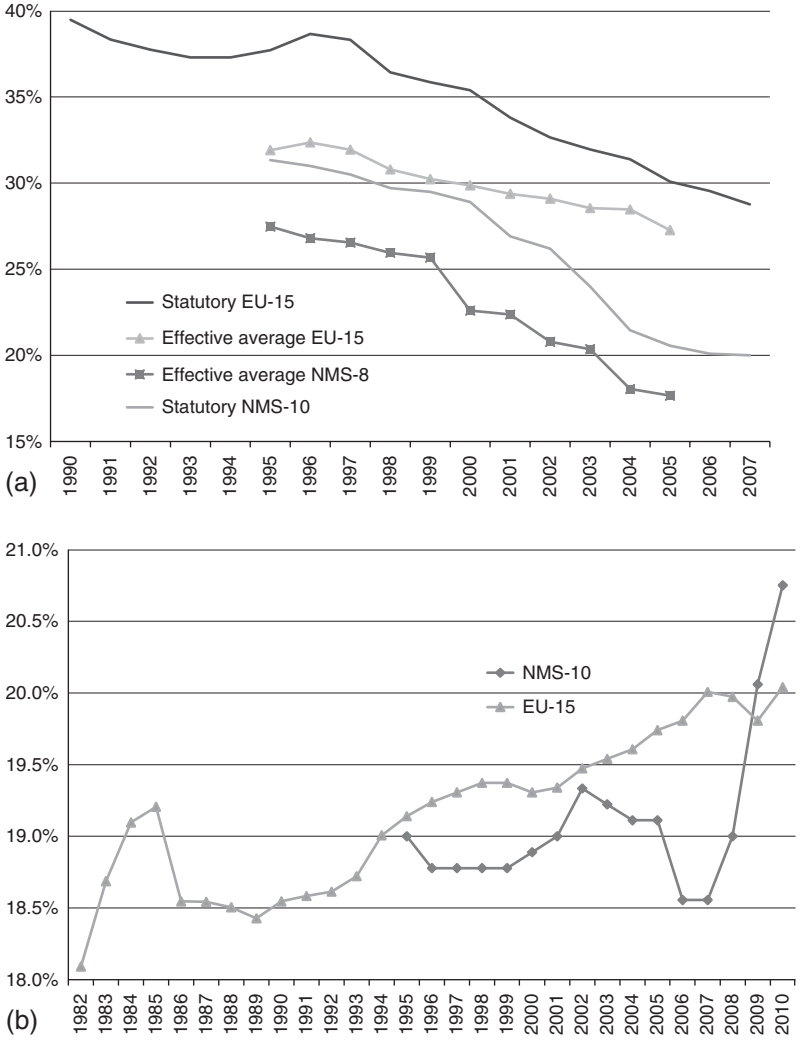


Figure 7.17 Taxing mobile and immobile tax bases in the EU. a) Corporate tax rate. b) Standard VAT rate in the EU.

Sources: a) Statutory rates, Devereux et al. (2002) and Eurostat and European Commission (2009); effective average tax rates, Overesch (2005). b) European Commission, DG Taxation and Custom Union, "VAT Rates Applied in the Member States of the European Union," taxud.d.1(2010) 118380.

Note: NMS-10, new member states of 2004 enlargement; NMS-8, NMS-10 minus Cyprus and Malta. Unweighted averages.

and markets are not too high (Andersson and Forslid, 2003). Government services themselves can contribute to this dynamics. The presence of firms in a location generates resources that can be used to provide local amenities such as infrastructure, education, etc., which, in turn, will attract new firms—the "bright light, big city" effect.

Empirical studies[29] confirm that location choices are primarily driven by the breadth of local demand. Cost factors, including taxation, also have an impact, but they are mainly secondary factors, especially in advanced economies. This means that tax differentials can be maintained under certain conditions: Higher taxes to the extent, for example, that they can be compensated for by geographic advantages or lower taxes to the extent that they compensate for geographic disadvantages. In the EU, for instance, this justifies higher tax rates in more-developed, large, core countries. In 2009, the average top statutory CIT rate was 23% in new member states and 24% in small, peripheral EU-15 countries, against 30% in large and/or central EU-15 countries.

Box 7.9 Tax Competition According to Zodrow and Mieszkowski[a]

Consider a representative household that consumes both a private good (in quantity x) and a public one g. The household's utility function is $U(x, g)$, with positive and decreasing marginal utilities in each of the two goods (and zero cross-utilities). The private good is produced by a representative firm through the following production function: $y = f(k)$, where k denotes the amount of productive capital, $f'(k) > 0$ and $f''(k) < 0$. In turn, the public good is delivered by the government by taxing private capital at a proportional rate t: $g = tk$. The public budget is balanced. The problem of the (benevolent) government is to set t so that private utility is maximized.

The household is endowed with a wealth K that is invested in shares of productive capital.

In a *closed economy*, the representative household holds the domestic capital stock: $K = k$. Its budget constraint is $x = f(k) - tk$. The first-order condition is written as:

$$u_g / u_x = 1 \qquad \text{(B7.9.1)}$$

where u_g, u_x denote the marginal utilities of the public and the private good, respectively.

In a *small, open economy*, the representative household chooses between domestic and foreign capital. The arbitrage condition is the equality

29. Surveyed by Hines (1999b, 2007), de Mooij and Ederveen (2001), Devereux and Griffith (2002), and Devereux (2006).

between the domestic, after-tax return and the world real interest rate, r: $f'(k) - t = r$. The first-order condition now is now written as:

$$\frac{u_g}{u_x} = 1 + \frac{\varepsilon_k}{1 - \varepsilon_k} > 1 \qquad \text{(B7.9.2)}$$

where ε_k is the elasticity of capital to the tax rate. Because u_g is a decreasing function of g, condition (B7.9.2) involves lower public good provision, hence lower equilibrium taxation than in the closed-economy case. An increase in taxation leads capital to flow out, which raises the pre-tax capital return until the after-tax return is back to the world interest rate. Due to the reduction in the stock of productive capital, domestic income falls; so does private welfare. This is why the equilibrium tax rate is lower than in a closed economy.

Allowing the public good to be productive or considering two large economies (that together determine the world interest rate) reduces the impact of capital mobility on the optimal tax rate, without canceling it. Finally, it can be shown that smaller economies are more prone to lower their tax rates, because the world capital return is more exogenous for them than for large economies.

[a]This box is derived from Krogsrup (2002).

b) Goods and services mobility

International trade provides a possible source of tax revenues, since, within the constraints of international trade agreements, a country can tax imported goods. As already mentioned, this is a major source of revenue for developing economies. However, just like other taxes, import and export duties introduce distortions that are costly in terms of social welfare (cf. box 7.3). The losers are uncoordinated consumers (consumption prices are higher due to the tariffs), while winners are organized, local producing firms. Unsurprisingly, political-economy models show that protection against imports is higher when producers are better organized and when price elasticities of foreign trade are lower (Grossman and Helpman, 1994). The difficulty of multilateral trade negotiations provides additional evidence of widespread resistance to tariff cuts and, more broadly, to international trade liberalization. However, a uniform tax on imports is equivalent (when associated with a uniform export subsidy) to a real depreciation of the domestic currency. Theoretically, in the long run it is absorbed by the adjustment of the real exchange rate. For example, if the trade balance is in surplus thanks to barriers on imports, the real exchange rate will appreciate. In practice, however, tax rates applied to various goods and services differ widely. They are generally much higher for agriculture than for manufactured goods, and among the latter, they can be very low—with exceptions (tariff peaks). Hence the macroeconomic logic does not really apply.

The development of intra-industry trade, i.e., trade of goods and services that are substitutable for each other (see chapter 6), implies that the demand for goods and services is very elastic to relative prices. Consumers are able to choose between goods and services coming from different countries. One consequence is that taxes on the production of goods are hardly passed on to consumers. For instance, a rise in social insurance contributions is unlikely to be passed on to consumer prices, because consumers can switch to foreign goods that do not suffer from the related cost increase. Similarly, efforts to reduce pollution by taxing emissions will be borne by those companies that are taxed, unless an international agreement is reached. Even a tax on the demand for goods may be passed on to the production side. Assume, for example, that the VAT rate is increased. In the short run, all goods that are consumed will be taxed the same way, be they imported or domestically produced. In the medium run, since the rise in VAT generates a loss in purchasing power for domestic workers, domestic wages may have to be increased, whereas this will not be the case in foreign firms. In the long run, even VAT would fall on domestic producers. There are many such issues related to the impact of taxes in an open economy. Studying them in more depth requires resort to international trade theory, which is not addressed in this book. One can refer to the textbook of Feenstra (2004).

7.3 Policies

As mentioned in section 7.1, tax policies aim at (i) collecting resources without introducing too many market distortions, (ii) redistributing incomes without discouraging labor supply or saving, and (iii) correcting specific market imperfections. The availability of a wide range of instruments makes these various objectives less contradictory than might appear at first glance. If taxation is so much debated, it is, of course, because of its impact on the various agents' disposable income, but also because tax incidence is generally poorly understood, because the agents' horizons may differ, because the model used to understand its effects can vary (perfect or imperfect competition, open or closed economy, etc.) and, of course, because of differences in the relative weights of the efficiency and redistribution objectives. Here we focus on how the theories presented in section 7.2 can be called on to address concrete tax policy issues.

7.3.1 Distributing the tax burden efficiently

Economic theory fails to provide any reliable tool for determining the optimum level of the total tax burden. As already mentioned, the Laffer curve does not provide useful guidance to policymakers in identifying this level: Absent tax collection problems, an economy as a whole generally lies on the left part of the curve, where a higher tax rate increases tax revenues whatever the initial level of tax pressure. Although high taxes mechanically translate

into large price distortions in the economy, Scandinavian countries seem to accommodate tax pressures as high as 50% of GDP. The choice of the general tax pressure is then left to social preferences, and especially to the desired generosity of the welfare state.

Theory is more prolix on how to distribute the tax burden in an efficient way, i.e., so as to raise taxes without introducing too many market distortions: Public revenues should be raised through low tax rates applied on large, relatively inelastic tax bases. However, several tax bases can be used: Consumption, payrolls, personal income, corporate income ... Which of them should be favored? It is safe to start from the long-run equivalence between social contributions, personal income taxation, and general consumption taxes.[30] With W denoting the nominal, unit labor cost for employers and Ω the purchasing power of employees, we have:

$$\Omega = \frac{(1 - t_{SSC2})(1 - t_{PIT})}{(1 + t_{SSC1})(1 + t_{VAT})} \frac{W}{P} \tag{7.2}$$

where t_{SSC1}, t_{SSC2}, t_{PIT}, t_{VAT} denote the rates of employers' social contributions, employees' social contributions, the personal income tax, and VAT (or any general consumption tax), respectively, and P represents the before-tax consumption price index. Equation (7.2) basically states that the four taxes have the same impact on workers' purchasing power. The ratio of W/P to Ω is called the *tax wedge**.[31] The distribution of these taxes between employers (who pay W/P in real terms) and employees (who receive Ω) only depends on the relative slopes of labor supply and labor demand, as detailed in section 7.2. If labor supply is steeper (less flexible) than labor demand, then W/P will remain unchanged whatever the taxes, and a tax increase will result in a fall in purchasing power Ω.

An important exception to this equivalence between taxes occurs at the minimum-wage level, because the latter is generally defined net of social contributions, but gross of VAT.[32] In this case, a rise in social insurance contributions mechanically raises the cost of labor W because the net wage received by employees cannot fall; in contrast, a rise of VAT causes a drop in the employees' purchasing power, if no compensation is made in the minimum wage. Therefore, policies aimed at encouraging the demand for low-skilled labor can use cuts in social contributions, because the latter result in lower labor costs while preserving the purchasing power (see figure 7.18). A similar policy at higher wages would result in an increase in wages and a constant labor cost, because the labor-supply curve is steeper at higher wages.

Another exception to tax equivalence occurs in the short run, before wage negotiations take place. In the short run, a rise in employers' social insurance

30. See, e.g., Malinvaud (1998), Sterdyniak et al. (1991).

31. For relatively high tax rates, $1/(1 + t) > 1 - t$, so t_{VAT} or t_{SSC1} have a slightly smaller impact on purchasing power.

32. As for personal income tax, households at the minimum wage are generally exempted.

Figure 7.18 Minimum wage and labor cost at the minimum wage in France (relative to median wage or labor cost).
Source: French Ministry of Economy and Finance.

contributions increases labor costs because wages cannot fall, whereas a rise in employees' social contributions, the personal income tax, or the VAT rate cuts the employees' purchasing power (since wages are not indexed in the short run). Hence, these various taxes have different stabilization properties. In 2007, Germany raised its standard VAT rate by three percentage points while cutting employers' social contributions by one percentage point. This tax package had a negative impact on consumption in the short run, due to its detrimental impact on purchasing power. Later in 2007, German unions asked for wage increases to compensate for the rise in VAT.

Finally, the above reasoning does not account for capital income, which is taxed through VAT and personal income tax (or withholding taxes), but is not subject to social insurance contributions, which are generally based on payrolls.[33] As a consequence, the VAT rate that is needed to raise a given amount of public revenue is theoretically smaller than the level of social insurance contributions that would be needed to raise the same revenue. In this case, VAT is to be preferred because it entails a lower rate. However, as detailed in box 7.10, the two tax bases are in fact similar because the share of capital income in GDP is close to the share of investment expenditures.

On the whole, although it may not rely on a larger base than labor taxes, VAT (or the sales tax) appears as the most neutral tax since, as shown in box 7.5, it affects labor and capital income in the same way and does not affect the consumption–savings trade-off. As shown in figure 7.5, general consumption taxes account for a growing share of tax revenues in OECD countries, from 13% on average in 1975 to 19% in 2007. Figure 7.19, however, shows that this proportion varies greatly across OECD countries, from more than 25% in Hungary or Iceland, to only 9% in Japan and 8% in the US.

33. This should be qualified, however. As suggested by the theory of tax incidence, capital income may de facto escape any form of taxation due to its high elasticity compared to other tax bases.

Box 7.10 VAT versus Social-Security Contributions

Consider a closed economy, where a tax is raised either on final consumption or on payrolls. In both cases, tax revenues are redistributed in a lump-sum way to households, there is no public consumption or investment, and no private investment. In this economy, final households' consumption C is equal to GDP Y:

$$Y = C \tag{B7.10.1}$$

Hence, the VAT base is Y. Suppose that the share of labor income in value added is 2/3, the last 1/3 being distributed to capital owners. The base of social insurance contributions is $2Y/3$. Suppose the government wants to raise revenues equivalent to $Y/5$. This can be achieved either through a 20% VAT rate or a 30% social insurance contribution.

Now introduce private investment I in this simple framework. We now have:

$$Y = C + I \tag{B7.10.2}$$

Since private investment is exempted from VAT, the VAT base is now narrower than Y. Assuming that $C/Y = 2/3$, the VAT rate that is necessary to raise a revenue of $Y/5$ is now 30%, the same as the rate needed to raise the same revenue through a social insurance contribution. In the golden-rule, long-run equilibrium and with a Cobb–Douglas production function (see chapter 6, section 6.2), the ratio of gross investment over GDP (I/Y) is equal to the share of capital income in GDP. In this case, the two taxes—VAT, social insurance contribution—rely on the same tax base.

Finally, let us introduce exports X and imports M into the analysis:

$$Y = C + I + X - QM \tag{B7.10.3}$$

where Q is the relative price of imports in terms of exports (equation (B7.10.3) is written in volumes of domestic goods). Since imports are taxed through VAT whereas exports are exempted, foreign trade raises the VAT base only to the extent that imports exceed exports, i.e., $X < QM$.

In the short run, substituting VAT for social contributions is likely to reduce imports because only the former weighs on imported goods. In the long run, however, higher imported prices are likely to push domestic wages upwards, so the advantage of VAT over social insurance contributions, for domestic producers, fades away, and imports are likely to be the same under the two types of taxes.

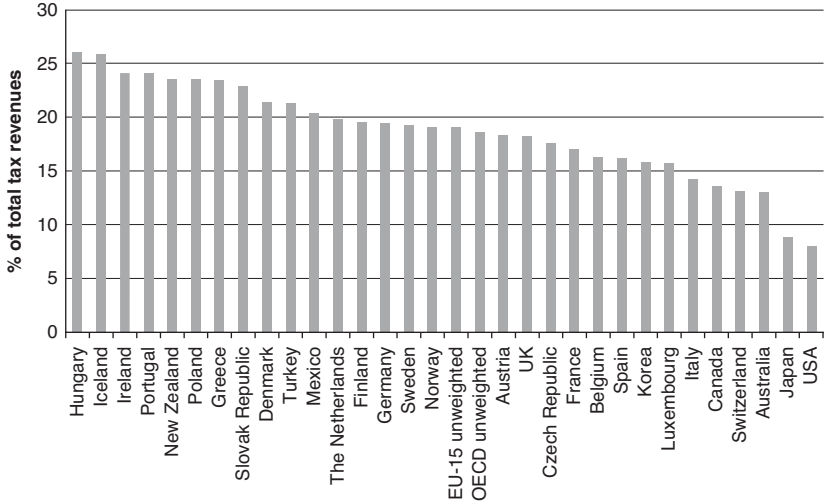


Figure 7.19 General consumption taxes in OECD countries, in 2007 (% of total tax revenues).
Source: OECD, *Revenue Statistics*, 2009.

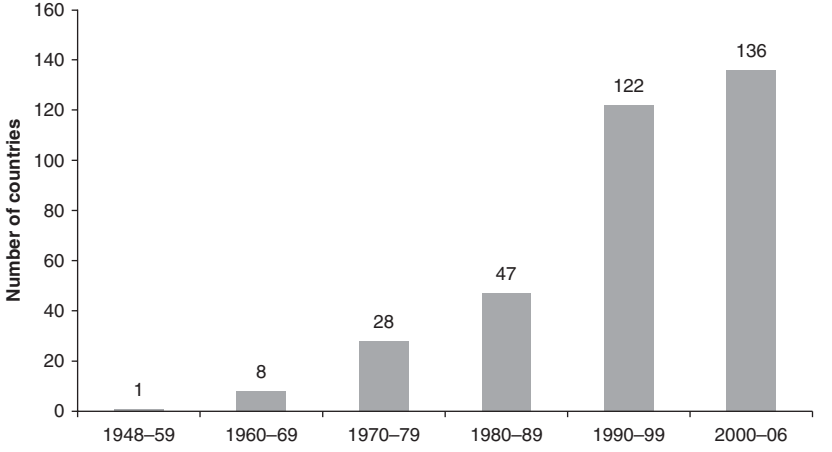


Figure 7.20 Number of countries applying VAT.
Source: US National Textile Association.

In most cases, the general consumption tax takes the form of a VAT. Indeed, VAT has progressively generalized to most countries of the world (figure 7.20). For instance, VAT is a prerequisite for EU membership. Some countries, like the US or India, do not use VAT but a Retail Sales Tax (RST). The latter is raised only on final consumption, whereas VAT is raised at each stage of the value-added chain (with appropriate tax credit for intermediate consumption). As detailed in box 7.11, VAT and the RST are equivalent from

an economic point of view, but VAT is generally viewed as more resistant to tax evasion.

Box 7.11 VAT versus Sales Tax[a]

Let us assume that a single producer of intermediate goods sells for 100 euros to a single producer of a final good that is sold 150 euros to final consumers.

- Under a 20% VAT rate, the producer of intermediate goods charges 20 euros (20% × 100) VAT to his or her customer and transfers this amount to the tax administration; the producer of the final good charges his/her customers 30 euros (20% × 150) VAT and transfers 10 euros (30–20, the VAT s/he already paid for the intermediate good) to the tax administration. The total tax revenue is therefore 30 euros.
- Under a 20% retail sales tax (RST), the producer of intermediate goods charges no tax and pays no amount to the tax administration. The final good producer charges 30 euros (20% ×150) to his/her customers, and transfers this amount to the tax administration.

The same rate of VAT and of RST produces, therefore, the same tax revenue. However, the VAT is generally viewed as preferable because it spreads the risk of noncompliance over a larger number of agents: If one firm within the supply chain fails to comply with VAT, the corresponding tax will be levied at the next stage of the chain; furthermore, suppliers have an incentive to register to charge the VAT since this will allow them to receive a refund for the VAT paid on the expenditure side; finally, there is no incentive under VAT to cheat on the nature of sales and declare a final sale as a business-to-business transaction.

Advocates of VAT argue that fraud is likely to be less widespread with VAT than with RST. However, VAT is also subject to fraud, through unreported sales, failure to register (small businesses), misclassification of commodities (when different rates are applicable), omission of self-deliveries, nonremittance of tax collected (for instance, through bankruptcy), imported goods not brought into tax. It is also subject to specific fraud mechanisms, such as false claims for credit or refund, fictitious "invoice mills" (companies that are set up solely to generate invoices that allow for VAT credit or refund, whether the corresponding VAT has been paid or not), "carousel fraud" (within the EU, where company A imports goods from another member state and sells the goods to company B; the latter is refunded for VAT paid on its domestic purchases, while company A did not pay any VAT on imported goods and disappears before the tax administration can ask it for VAT received on its sales to B).

Within the European Single Market, fraud on VAT is thought to be as large as 10% of net VAT receipts in some member states, see Mathis (2004). A way to reduce it is to avoid multiple VAT rates or to refund VAT relatively slowly (so that firms always remain net creditors to the tax administration). On the whole, VAT seems to be more resistant to fraud than the RST. It is a fact that VAT countries have been able to apply high rates (25% in Northern Europe) whereas RST countries have generally applied rates below 10%.

[a]This box draws on Keen and Smith (2006).

Most developing countries have also adopted VAT systems. However, some sectors (services, wholesale and retail sectors) are often left out, which significantly reduces VAT receipts. In some countries, creative schemes have been introduced to raise the incentive for the final consumer to pay the VAT, for instance through refund systems. In China, local tax authorities introduced in many provinces a system through which retailer VAT receipts are coupled with lottery tickets.[34] Enforcing VAT in developing countries has become a crucial issue especially since import tariffs are to be cut as a result of WTO membership.

In contrast, VAT is generally not a good instrument for targeted actions, such as promoting employment in selected sectors. Consequently, the 1999 EU directive that experimentally allowed reduced VAT rates on a list of labor-intensive services (see box 7.17) is questionable. Even if VAT cuts are passed on to consumption prices, and if demand reacts positively to the price cut, the impact of the VAT cut is likely to be diluted across the various production factors (low-skilled labor, skilled labor, capital, intermediate consumption). To encourage demand for low-skilled labor, it is therefore preferable to act directly on labor costs through cuts in social insurance contributions.

Although VAT is an efficient way to raise public revenues, the budget of a country cannot rely solely on it, for various reasons:

- VAT is not a progressive tax and social preferences may call for some redistribution through the tax system;
- In line with Olson's (1969) equivalence principle, it is desirable to finance local government services through local taxes (see chapter 2);
- The tax system may be used to modify relative prices on purpose, for instance to correct market imperfections.

The question arises as to how to introduce progressive taxes in the most efficient way, i.e., without introducing too many distortions. This is, in fact, a complicated task. For instance, the existence of a personal income tax calls

34. See Marchese (2007).

for a corporate income tax (otherwise, there is an incentive for households to "incorporate" so as to declare their revenues as corporate income). In turn, corporate tax introduces three distortions. First, it raises the cost of investment, thus reducing capital accumulation; second, it changes the relative cost of investment depending on the way it is financed, since interest payments are generally deductible from taxable profits whereas dividends are not; finally, it introduces distortions between multinationals, affiliates, and local firms that may not receive equal tax treatment (see box 7.12).

Box 7.12 The Search for an Efficient Corporate Tax[a]

The relevance of taxing corporate income may be questioned since capital is internationally mobile and the burden of a corporate income tax (CIT) can thus be shifted to immobile tax bases. The corresponding income may be taxed at the less-mobile shareholder level, i.e., as a personal capital income. Still, the CIT may be justified on several grounds: (i) The corresponding income may be easier to trace at the corporate level than at the individual one; (ii) the tax base is easier to measure at the corporate level, especially when taxing rents rather than total profit is at stake;[b] (iii) the CIT may be used as an (imperfect) substitute for missing fees for the use of government services by corporations; (iv) the CIT is the only way to tax foreign capital owners; (v) the CIT acts as a backstop for the personal income tax; and (vi) from a political point of view, it may be less difficult to directly tax corporations than actual voters—workers or capital-owners.

Still, taxing corporate income raises two series of questions:

- Should the tax be raised in the country where the activity takes place (*source principle**), in the country where capital-owners (either individuals, headquarters or institutional investors) are located (*residence principle**) or in the country where the goods and services are finally consumed (*destination principle**)?
- Should the tax fall on the full return on equity (including both normal return and rents), on the full return on capital (including debt-financed capital) or only on rents (excluding "normal" return by exempting interest payments and "normal" dividends)?

In most countries, the CIT is raised under the source principle and repatriated profits from foreign affiliates are exempted from any taxation. However, in some countries including the US and the UK, there is a credit for taxes paid abroad on affiliates' profits, so that the residence principle de facto applies to the multinationals headquartered in the country. Due to this tax credit scheme, foreign affiliates of, say, UK multinationals do not receive the same tax treatment as local firms abroad, or as affiliates of multinationals headquartered in exemption countries.

In most countries, the tax base is the full return on equity. In particular, interest payments are deductible from taxable profit. This tax base, combined with the source principle, is especially vulnerable to tax optimization by multinationals. Indeed, multinationals can shift profit from one country to another through *transfer pricing** (e.g., over-pricing intermediate goods sold by those affiliates located in low-tax countries) and intra-firm finance (e.g., loans from affiliates located in low-tax countries to those located in high-tax ones), which has led governments and international organizations (especially the OECD) to try to impose codes of conduct.

To remove some of these distortions, several tax reforms have been proposed and sometimes adopted. One of them would aim at taxing only rents, not the "normal return," by introducing an allowance for the cost of equity finance. Such a CIT system, which was applied in Croatia from 1994 to 2001, used in Brazil, and introduced in Belgium in 2006,[c] reduces the distortions related to the CIT, since debt and equity finance are treated equally only extra profits ("rents") are taxed, and tax optimization through intra-firm loans is reduced. The main disadvantage of this system is that it amounts to narrowing the tax base, which leads to a rise in the statutory tax rate if constant receipts are needed. This re-introduces the risk of more distortions as well as of tax evasion.

Another proposal consists, on the contrary, in removing interest payment deductibility. Again, debt and equity finance would be treated equally so that both would be taxed, which means a broadening of the tax base and a lower statutory tax rate. Such a *Comprehensive Business Income Tax** (CBIT) was proposed by the US Treasury in 1992.[d]

[a] This box relies on Devereux and Sørensen (2006) and Auerbach et al. (2007).
[b] The taxation of economic rents is theoretically nondistortionary, since the normal return generated by the marginal investment project is exempted.
[c] The interest deduction for risk capital (a.k.a. *Notional Interest Deduction**) was introduced in Belgium in 2006 to replace the special tax regime for "coordination centers"—a system that was banned as discriminatory by the European code of conduct. The interest deduction is calculated as a notional interest rate (itself a moving average of past 10-year interest rates on Government bonds) multiplied by the company's equity.
[d] In fact, the initial proposal consisted in a relatively high statutory tax rate but a reduction in personal capital-income taxes. See Auerbach et al. (2007).

Furthermore, corporate income tax introduces a distortion at the household level between labor (or noncorporate capital) income and corporate-capital income, since the latter is taxed twice (at the corporate level, and then at the personal level). Then, an imputation system or a dual-rate system needs to be introduced in personal income tax.[35] These various corrections result in rather complex tax systems (box 7.13).

35. For instance, Scandinavian countries use a dual-rate system in which capital income is taxed at a low, flat rate, whereas progressive taxation is applied on labor income.

Box 7.13 Tax Complexity

The search for efficiency and equity often leads governments to introduce special schemes to correct a specific inequality (for instance, accounting for the number of children) or encourage specific behaviors (e.g., tax allowances for health care or college expenditures, or contributions in pension funds). These schemes accumulate over time, adding to the complexity of the tax code. In turn, this complexity reduces the efficiency of the system because higher marginal tax rates are needed to raise a given tax revenue. It also undermines the social acceptance of taxes.

The cost of tax complexity can be borne either by taxpayers or by tax administrations, or both. In the US, compliance costs borne by taxpayers are estimated at around 22 cents for every dollar collected (Hodge et al., 2005). Due to the alternative minimum tax scheme, households are required to calculate their tax liability under two alternative systems (with different tax breaks, etc.), before selecting the one leading to the highest liability. As for businesses, they face 17 different categories of interest expenses (Graetz, 2007).

As for tax administrations, collection costs are estimated at around 1% of net tax revenues (from 0.5% to 1.7%, see figure B7.13.1). Reducing collection costs is a major challenge for tax policy.

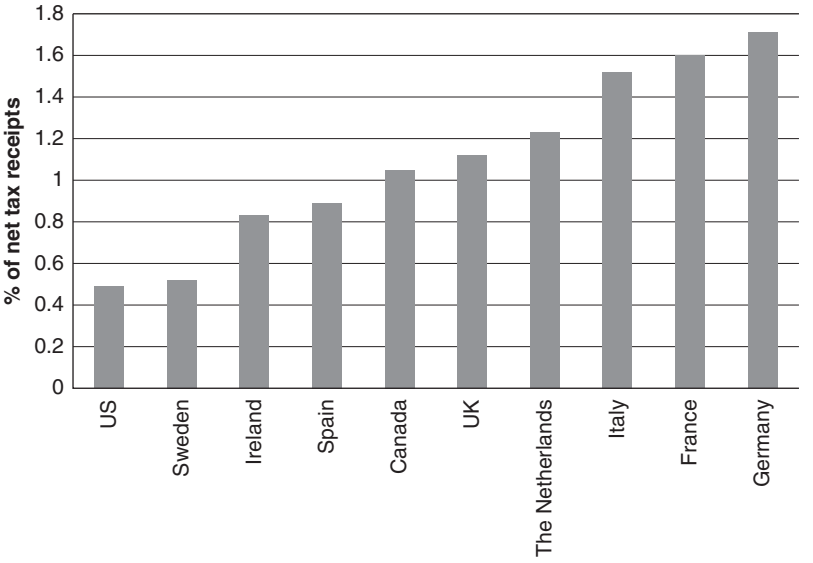


Figure B7.13.1 Tax administration costs in percentage of net tax receipts, 1997. Source: Lépine (1999).

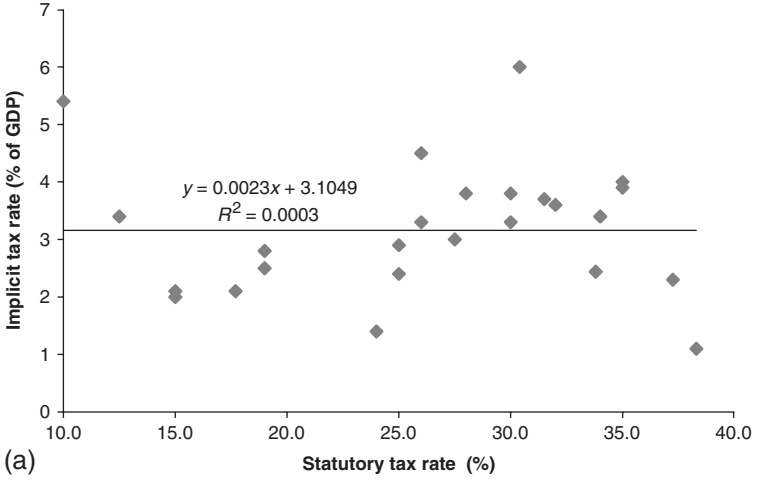
Finally, the mobility of tax bases has strong implications for the design of an efficient tax system. Indeed, when the tax base is elastic to the tax rate, which happens if tax bases are mobile internationally, taxes appear more distortionary (see section 7.2).

At the local level, the mobility of firms and workers limits the scope of taxation. This is especially the case in federal countries where states (or Länder, or cantons) have wide tax autonomy. In Switzerland, for instance, high-income individuals and companies tend to locate in relatively low-tax cantons. This mobility of taxpayers has triggered a race-to-the-bottom between some cantons (Obswald, Zug, Appenzell) that strive to attract wealthy households through tax cuts, sometimes turning personal income tax into a regressive tax.[36] Corporate income tax is also submitted to tough competition, for instance in Germany where each commune fixes freely the rate of local corporate taxation (*Gewerbesteuer*). In contrast, local tax competition is limited in the UK where a low proportion of local expenditures are financed through local taxes, and where companies are taxed at the same rate everywhere (*Uniform Business Rate*).

National tax systems also face international competition. Until the early 2000s, tax competition was limited within the EU. Admittedly, statutory corporate income tax (CIT) rates had declined during the 1990s, but this trend had been compensated by a broadening of the base (Devereux et al., 2002). However, sharp cuts in statutory CIT rates initiated in the early 2000s were not fully compensated by base broadening, and effective tax rates began to decline. Tax competition should theoretically be limited by the need to raise revenues: A race-to-the-bottom is unlikely if corporate tax revenues, which account for around 10% of tax receipts in the OECD (see figure 7.5) are needed in order to finance government services. The problem with this line of reasoning is that there is little connection between corporate tax rates and corporate tax receipts, as evidenced in figure 7.21 for EU countries. This disconnect can be interpreted as another illustration of the Laffer curve: A higher tax rate does not necessarily increase tax revenues because the tax base contracts (or, for that matter, relocates in another constituency). All countries could be expected to converge to the revenue-maximizing rate—estimated at around 23–33% by Clausing (2007) or Devereux (2007)—but they do not. Many countries have cut their CIT rates to less than 20%.

An alternative interpretation of figure 7.21 is country heterogeneity. In particular, the EU is composed of countries of unequal size, which will face different elasticities of the tax base to the tax rate. For some countries, cutting CIT rates may result in higher tax receipts, whereas this is not the case for larger countries. However, tax cuts on mobile bases are obviously noncooperative strategies, and it is far from granted that a race-to-the-bottom of, say, new

36. See Brülhart, M., *Le Temps*, March 14, 2006. Kirchgässer and Pommerehne (1996) and Feld and Reulier (2008) show that cantons set their tax rates strategically, depending on other cantons' tax rates.



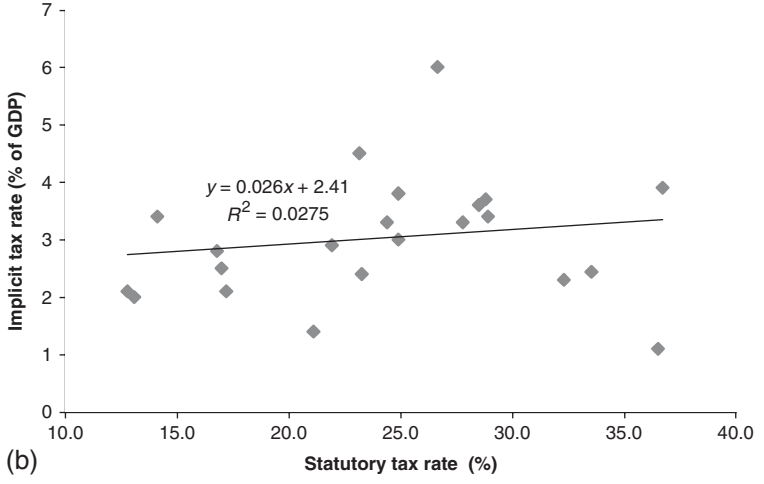


Figure 7.21 Corporate tax rates and revenues. a) Statutory and implicit CIT rates in 2005; b) EATR and implicit CIT rates in 2005.
Sources: Statutory rates, Eurostat and European Commission (2007); effective average tax rates, Overesch (2005).

Note: EATR stands for Effective Average Tax Rate. It corresponds to the average tax rate of a unit investment with average pre-tax return. It is calculated based on tax codes and a number of assumptions concerning the type of investment, the way it is financed, its return, etc. The implicit CIT rate is calculated as the ratio of CIT revenues to GDP.

member states, will succeed in securing foreign-investments-driven catch-up such as Ireland achieved with its 12.5% statutory rate.

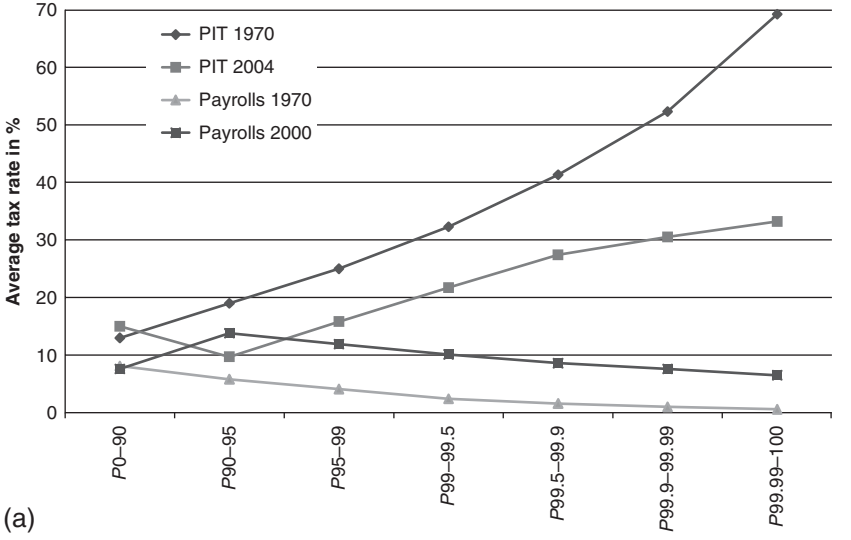
International tax competition also affects the personal income tax (PIT): PIT rates on the highest income brackets have declined over the past 20 years in OECD countries. On the top of this movement, a number of countries have introduced special tax regimes for "impatriates," those foreign, high-level executives that temporarily work in one country as employees in affiliates of multinational companies. Lower marginal rates for impatriates are designed to encourage inflow of high-level workers, which are viewed as complements to foreign direct investments. They are consistent with the Ramsey rule, which suggests that highly elastic tax bases should be taxed less. However, they introduce a distortion between local workers and foreign ones, since the marginal tax rate is different for both populations, which may reduce the incentive of the domestic population to invest in human capital. More importantly, impatriate regimes obviously increase inequalities across households, while the PIT is supposed to be used to reduce inequalities.

7.3.2 Distributing the tax burden equitably

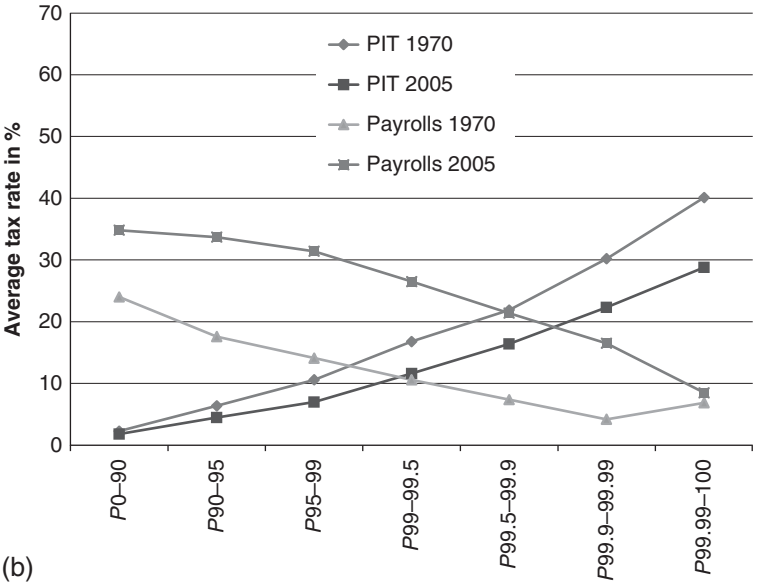
The principles of burden-sharing mentioned in the previous section, that portrays general taxes on consumption as the most neutral way of raising public revenues, are only concerned with economic efficiency. Switching to redistribution or, more modestly, to equity concerns, general consumption taxes are no longer the appropriate instrument since they are indirect taxes that apply proportionally, whatever the consumer's income. Conversely, personal income tax (PIT) makes it possible to perform interpersonal redistribution through progressive taxation. In fact, in OECD countries, the PIT on average represents a higher share of public revenues than VAT (25% for PIT against 19% for VAT in OECD countries in 2007, see figure 7.5). Progressive taxation targets *vertical equity**, as opposed to *horizontal equity**, which aims at an equal treatment of various forms of income.

a) Vertical equity

PIT and wealth taxes are the traditional instruments of income redistribution through the tax system. For example, figure 7.22 shows that the PIT average rate is higher for higher incomes in the UK, France, and the US, although the degree of progressiveness fell dramatically between 1970 and 2000 in the UK. However, payroll taxes are shown to be regressive in the same figure, i.e., the average tax rate is lower for higher incomes. Hence the whole tax system needs to be considered when measuring the progressivity of taxes in any given country. In France, for instance, the large increase of regressive payroll taxes between 1970 and 2005 reduced the overall progressiveness of the tax system (see Piketty and Saez (2007)). For the US, Piketty and Saez find

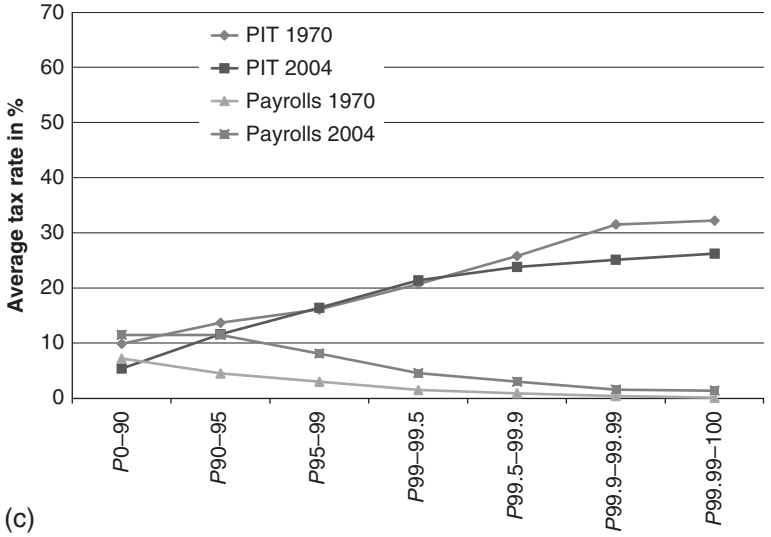


(a)



(b)

Figure 7.22 Average tax rates for the personal income tax and payroll taxes for various percentiles of the population. a) UK, b) France.



(c)

Figure 7.22 (Cont'd) c) US.

Source: Piketty and Saez (2007).
Note: $Px-y$ stands for the percentile between x% and y% of the population.

that the progressivity of the tax system has been substantially reduced since 1960 by the marked fall in the corporate tax rate.[37]

The redistribution motive raises the traditional trade-off between efficiency and equity. As underlined by the optimum taxation theory, distortions induced by progressive income taxation carry an economic cost. In some cases, it is possible to raise both efficiency and equity by redesigning a tax. For instance, when a high statutory tax rate is levied on a narrow tax base (due to the multiplicity of exemptions), it can be more efficient and more equitable to reduce the rate while broadening the base. This was the case with the US tax reform in 1986, which combined personal income tax rate cuts and base-broadening. After the reform, fewer individuals were able to escape the tax, whereas those who already complied with it benefited from lower rates. However, most tax reforms of the 1990s and 2000s have resulted in a flattening of the marginal tax rate curve, with rates falling more markedly for high-income brackets. The most radical examples are provided by countries having implemented *flat tax** systems, i.e., tax systems with constant marginal tax rates (table 7.1). Slovakia was, in 2005, the first OECD country to introduce a flat rate personal income tax, in the context of a sweeping reform of its tax system that also unified (at 19%) the PIT, CIT, and VAT rates (Brook and Leibfritz, 2005). In theory, a flat tax system may achieve both efficiency and redistribution when combined with a generous basic allowance (a fixed

37. Piketty and Saez (2007) assume that corporate taxation falls entirely on capital income.

Table 7.1
Countries with a flat tax system

	Flat tax adopted	Personal income tax rates		Corporate income tax rate, after reform	Change in basic allowance
		After	Before		
Estonia	1994	26	16–33	26	Modest increase
Lithuania	1994	33	18–33	29	Substantial increase
Latvia	1997	25	25 and 10	25	Slight reduction
Russia	2001	13	12–30	37	Modest increase
Ukraine	2004	13	10–40	25	Increase
Slovak Republic	2005	19	10–38	19	Substantial increase
Georgia	2005	12	12–20	20	Eliminated
Romania	2005	16	18–40	16	Increase

Note: Most countries do not apply pure flat tax systems since the flat rate does not apply to all tax bases. For instance, social insurance contributions are levied separately. To the extent that these contributions exceed the present value of future (contingent) social benefits, the system is not neutral, since labor and capital income are not taxed equally. See OECD (2006).

Source: Keen et al. (2007).

income level that is not taxed, see section 7.1). Taxpayers are exempted on their first units of income. In practice, however, a flat tax system generally leads to much flatter average tax rates, as illustrated in box 7.14 in the case of Russia.

Box 7.14 The Russian Flat Tax

In 2001, a flat tax system was introduced in Russia. Before 2001, a household earning less than 3168 rubles was exempted from the personal income tax. Higher incomes were then taxed at three rising tax rates corresponding to successive income brackets: 12%, 20%, and 30% (see figure B7.14.1).

The reform of 2001 increased the basic allowance to 4800 rubles but reduced the marginal tax rate to a flat, 13% rate. As illustrated in the figure, the new PIT remained a progressive tax, but essentially at the lower end of the income scale. Above 30000 rubles, the progressiveness almost disappeared: With the removal of tax brackets, the average tax rate converges rapidly towards the flat, low marginal tax rate. At that time, however, Russia was plagued with a very low level of tax compliance. With low marginal rates, high-income households became less reluctant to pay taxes. Hence the final outcome of the reform was less regressive than it appears at first sight, since a number of wealthy households started to pay taxes (see Ivanova et al., 2005).

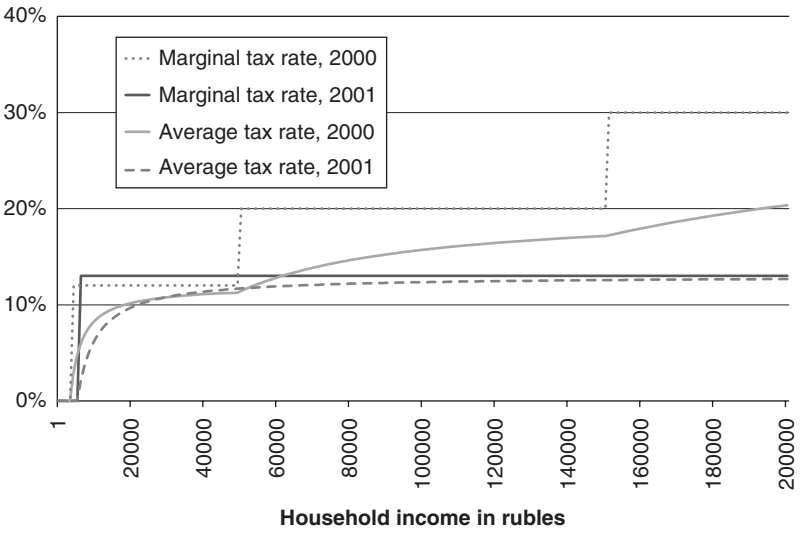


Figure B7.14.1 Marginal and average personal income tax rate in Russia. Source: Ivanova et al. (2005), authors' calculations.

In the US, the fall in marginal tax rates that started in the early 1980s and continued with the fiscal reforms of 1986 and of the early 2000s, has benefited those on high incomes much more than those on modest ones (the reform of 1986 being an exception, cf. supra). Furthermore, these tax cuts were not matched by corresponding spending cuts. Depending on the way cumulated public deficits are to be financed in the future, the overall diagnosis on the redistributive impact of tax reforms can be dramatically altered (see box 7.15 for the specific case of US tax reforms carried out between 2001 and 2006).

Box 7.15 The Redistributive Impact of the 2001–06 US Tax Cuts

From 2001 to 2006, the US adopted a new tax reform almost every year: Economic Growth and Tax Relief Reconciliation Act (2001), Jobs and Growth Tax Relief Reconciliation Act (2003), Working Families Tax Relief Act (2004), Tax Increase Prevention Reconciliation Act (2005), Pension Protection Act (2006). Taken together, these various reforms amounted to a tax cut of approximately $2 trillion until 2010 (when the measures were initially scheduled to expire). Since these tax cuts concerned mainly

capital income and personal income taxes, they benefited all taxpayers, but more particularly the higher percentiles of the population (see figure B7.15.1).

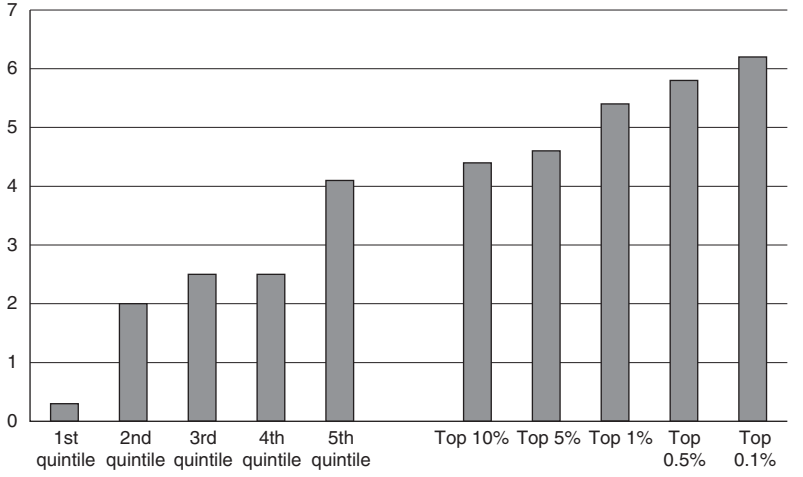


Figure B7.15.1 The direct impact of US 2001–06 tax cuts on income distribution: Percentage change of after-tax income in 2007.
Source: Leiserson and Rohaly (2006).

However, these tax cuts were not financed and the budget deficit increased. To the extent that tax cuts are made permanent, they must be financed through either expenditure cuts, or tax increases. Figure B7.15.2 reports the distributional effects of the 2001–06 tax cuts calculated by Leiserson and Rohaly (2006) under three alternative assumptions concerning their financing: (i) A lump-sum tax (or, equivalently, a cut in public expenditures that affects all citizens equally), (ii) an additional tax proportional to income, and (iii) an additional tax proportional to the tax liability. Not surprisingly, lump-sum financing accentuates the regressiveness of the 2001–06 tax reforms, since its combination with the tax reforms leads to a fall in the after-tax income of the first quintile by 18.5% in 2010, and to a rise in that of the highest quintile by 2.6%. In contrast, financing proportional to the tax liability almost turns the 2001–06 reform package into a neutral package, with smaller percentage changes in after-tax incomes across the various percentiles of the population (although there are substantial differences within the top quintile).

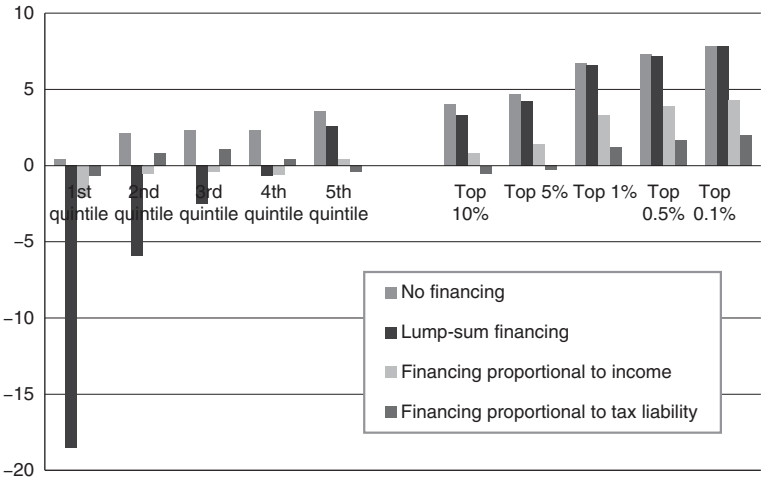


Figure B7.15.2 Redistributive effects of US 2001–06 tax cuts when accounting for their financing: Percentage change in after-tax income in 2010, for various percentiles of the population.
Source: Leiserson and Rohaly (2006).

During the 1990s, it was suggested that redistribution could be reconciled with efficiency by applying a constant marginal tax rate coupled with a *universal transfer**, i.e., a transfer that would be given to any individual or household, regardless of their income. Coupled with a constant marginal tax rate, a universal transfer could theoretically achieve the same level of redistribution as a progressive marginal rate system: Under a certain threshold, taxes paid would be over-compensated by the universal transfer; in the middle of the income scale, taxes paid and the universal transfer received would be about the same; finally, at the higher end of the income scale, taxes paid would exceed the universal transfer.[38] In practice, however, it is generally considered that, in order to maintain an identical redistribution level, the constant marginal tax rate would need to be very high (cf. box 7.16). Hence, no

38. The idea of a universal transfer paid to any individual independently of his/her income goes back to the sixteenth century when it is said to have been (unsuccessfully) proposed to the mayor of Bruges. At the beginning of the nineteenth century, the *Speenhamland system* was the first natural—and unhappy—experience of universal allocation: In Speenhamland, a district located in southern England, magistrates decided that the parish would supplement peasants' income up to a certain subsistence level, based on the price of bread and the number of children. This system spread quickly in the south of England. But Thomas Malthus criticized this encouragement to have children without being able to provide for their needs. The idea, nevertheless, was taken up by the utopians of the late nineteenth century, then again in the 1930s and 1940s in the UK by the economist James Meade, and finally in the twentieth century by Lady Juliet Rhys-Williams, who proposed this system as an alternative to the Beveridge report.

country so far has introduced such radical reform, even if some have applied schemes that are close to negative taxation (cf. box 7.1).

Box 7.16 Universal Transfer cum Flat Tax

The idea is to provide any citizen with a guaranteed lump-sum income. As a counterpart, households are to pay taxes on the very first unit of income, at a constant marginal rate (see figure B7.16.1). For poor households, the tax due is lower than the lump-sum transfer; thus, they receive a net transfer from the tax administration, which amounts to negative taxation. For the others, the universal transfer is deducted from the tax due.

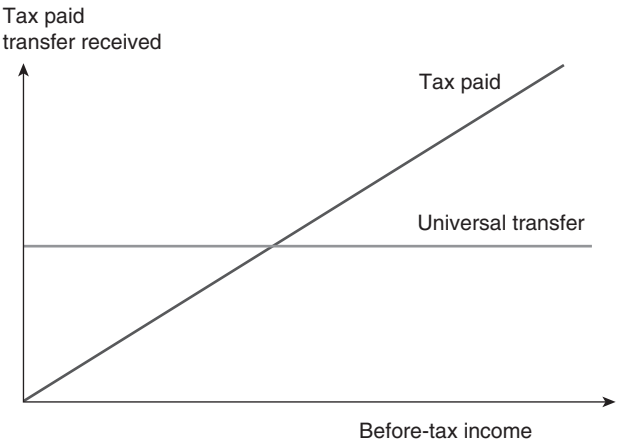


Figure B7.16.1 A constant marginal tax rate coupled with a universal transfer.

The advantage of this system is that the marginal tax rate is constant, which minimizes any disincentive to work. However, this type of system supposes a relatively low universal transfer, which is not very redistributive. Indeed, the higher the universal transfer, the higher the (flat) tax rate needs to be (to finance the transfer). In the case of France, Bourguignon and Chiappori (1998) found that the tax rate needed to finance a universal transfer of 4500 euros a year per "equivalent single adult" was 35% of households' income, a rate that corresponds to a high PIT bracket marginal rate. The choice between a flat tax cum universal transfer and a progressive marginal tax system then depends on social preferences—see Atkinson (1970).

Equity concerns also have a dynamic, intergenerational dimension. Indeed, progressive income taxation limits wealth accumulation. Combined with inheritance taxes, it reduces inherited inequalities across individuals while also

limiting individual incentives to climb the social ladder. In the case of France, Piketty (2001a, b) showed that this mechanism has been powerful in limiting the increase of inequalities. The problem arises again if wealthy households can escape taxation, for instance, by relocating their income. Since the 1990s, a Laffer curve may have appeared due to higher capital mobility for the highest percentile of the population. If such is the case, taxing the highest percentile households at a lower rate would raise rather than reduce tax receipts from them, due to a lower incentive to escape taxation.

b) Horizontal equity

*Horizontal equity** consists in ensuring that any income is taxed equally, whatever its origin. This raises the difficult question of how to tax capital income.

Taxes on dividends raise a specific issue since the dividends received by the shareholders have already been taxed through the corporate income tax. This double tax on dividends obviously contributes to making the fiscal system more redistributive. However, it introduces horizontal inequality between those households that invest their savings in equity, and those who choose to hold bonds or real estate. It also creates a corporate-finance distortion, since companies will prefer to finance real investment through bank loans or bonds (interest payments are deductible from taxable profits) rather than through equity or retained earnings. This distortion can induce *thin capitalization**, i.e., insufficient capital or too much debt. Therefore, dividends received are sometimes exempted from the PIT (e.g., Greece), half-exempted (e.g., Germany, Austria, Belgium, Luxembourg) or taxed at a low, flat rate (e.g., Ireland, Denmark, Portugal, the UK, the US). Furthermore, capital gains are exempted from taxation in a number of countries, including Germany, Austria, Belgium, and Luxembourg. This system raises a vertical-equity problem insofar as capital income, which is concentrated on the wealthiest households, tends to escape tax progressiveness. In some countries (Spain, Finland, France, Sweden), the income tax is calculated on all incomes received, but taxes already paid as corporate taxes are deducted from the tax invoice (this system is called a *tax credit**, as opposed to full or partial exemption of dividends in the other cases mentioned).

These exemption or credit systems that aim at eliminating (or reducing) double taxation on dividends however introduce two additional distortions. The first one relates to the nationality of firms. Exemption or credit does not generally apply to dividends received from foreign companies, which induces savers to hold domestic rather than foreign shares and may impede portfolio diversification.[39] The second distortion stems from the differentiated treatment of capital income (taxed at a reduced rate) and labor income (taxed

39. Joumard (2002). From 2004, the European Commission has asked member states to eliminate such distortions, which constitute impediments to the European Single Market.

at the PIT rates), which may distort capital accumulation towards financial rather than human capital.

Beyond these examples, tax systems have accumulated many minor distortions over time, often due to exemptions prompted by lobbies. The notion of *tax expenditure** refers to the loss in tax revenues related to all these specific measures; it can be interpreted as equivalent to subsidies in favor of various interest groups (see chapter 2 for a discussion of the political economy of interest politics). Tax expenditure may be as high as 20% of tax revenues in France, 40% in Spain, and 50% in the US.[40] In democracies, the call for simplification and base broadening (through reduced exemptions) of the tax system is a leitmotiv of newly elected governments, but generally does not resist the pressure from lobbies and the pursuit of reelection.

7.3.3 Correcting market failures

As illustrated above, it is difficult in practice to design a tax system that would reconcile neutrality and redistribution. However, in some cases, neutrality is not searched for, just the opposite, as the tax is then designed to correct market imperfections. Introducing nonneutral taxes may contribute to making markets more efficient.

a) Paternalist taxation

What is the difference between a social insurance contribution and a private contribution to a pension fund or a health insurance scheme? Both aim at protecting the individual against the risk of getting old without resources or of having to support costly medical care. The only difference is the compulsory nature of social contributions, as opposed to free contributions to private schemes (and the choice between various schemes). Why, then, impose a public social insurance system financed through taxation? Two reasons may be put forward: Equity and individual myopia (or lack of rationality).

- *Equity:* A compulsory system allows for redistribution across individuals. For instance, the cost of a given illness is basically the same whether the patient is rich or poor. Leaving each individual responsible for his/her own insurance through the private system is therefore anti-redistributive. Some households may not be able to pay for the insurance scheme. Having a single, compulsory system allows for cross-subsidization from the richer contributors to the poorer. It can also create incentives for the poorer to take costly preventive measures such as vaccination and consulting a doctor when sick, which has positive externalities on other individuals (lower risk of contagion) and on public finance (lower pressure on public-funded hospitals).

40. Conseil National des Impôts (2003).

- *Myopia or lack of rationality:* If individuals are myopic, then they may not correctly insure against the various risks they incur. For instance, they may be over-optimistic concerning their ability to work during their old age or not well-informed about their life expectancy. Also, *financial illiteracy** is well-documented: Most households do not master the basics of risk, return, and portfolio choice (Lusardi and Mitchell, 2007). Besides, it is well-known from behavioral economics that people do not exhibit rational expectations, and that they exhibit dynamic inconsistency (see, e.g. Kahneman and Tversky, 2000, and chapter 2 of this book). For these reasons, the government may wish to force, or at least incite, individuals to hedge against some risks. The same idea can justify policies aimed at encouraging households to save, for instance, through a tax exemption on voluntary contributions into pension funds or in some popular savings vehicles (life insurance . . .) or through owning their house (exemption of mortgage interest payments).

These targeted tax exemptions, however, introduce distortions insofar as they modify the relative yield of the various savings vehicles, for instance, between investing in housing, bonds, or equity. Hence, a specific distortion (excessive preference for the present) is replaced by another one (distortion across savings vehicles).

In advanced economies, tobacco and alcohol are heavily taxed, notably on grounds of public health since individuals may not properly assess the risks involved in consuming too much of these items. Taxes are designed to make individual behaviors fit a "safe" behavior defined by the government.[41] In the same vein, taxes have been proposed on sodas or fatty food in order to fight child obesity. Opponents of "fat taxes" argue that it is anti-redistributive; since sodas and fatty items are cheaper than healthy fruits and vegetables, they are consumed in larger quantities by the less wealthy households. Another way to encourage poor households to consume healthier food could be to reduce taxes on agricultural imports, because this would lower the consumer prices of farm products. More radically, some economists consider paternalism to be contrary to the freedom of choice, which is at the heart of free markets. Milton Friedman was the herald of this approach, as exemplified in the case of Social Security, in the following judgment:

> I believe that it is not the business of government to tell people what fraction of their incomes they should devote to providing for their own or someone else's old age.
>
> Milton Friedman, "Social Security: The General and the Personal," *Wall Street Journal*, March 15, 1988

41. The corresponding tax revenues are, of course, welcomed by the government. However, it should be noted that there is some contradiction between using such a tax to curb private consumption (which relies on high elasticity of consumption) and the wish to raise public revenue (which necessitates low elasticity).

It can be objected that paternalism does not go against individual freedom as long as it does not involve coercion (Thaler and Sunstein, 2003). Tax policy is well suited for this purpose, as long as the tax level is not confiscatory.

b) Environmental taxes

Whereas paternalist taxation aims at responding to households' lack of information or to their too-short horizons, environmental taxes implement the polluter-payer principle and aim to have polluters internalize the externalities they produce, along the lines of box 7.8. *Energy taxes**, which mainly aim at raising public revenue from a relatively less-elastic demand, need to be distinguished from *environmental taxes**, or *green taxes**, intended to curb the behavior of taxpayers. The former have traditionally been much higher than the latter. In 2007, energy taxes represented 1.8% of GDP in the EU-27, contrasting with 0.6% of GDP for transport taxes and 0.1% of GDP for taxes on pollution and on the use of natural resources (source: European Commission, 2009). On the whole, energy and environmental taxes still represent a very small share of compulsory levies in OECD countries (see figure 7.5 in section 7.1).

Denmark provides an interesting case study. A tax on energy was introduced after the first oil shock and it applies today to any energy source in that country. Like VAT, this tax is recovered when it is paid on intermediate consumption. In 1991, a tax on carbon dioxide emissions was introduced at a high level (13 euros per ton of CO_2), but with partial exemptions for energy-intensive companies. This tax was not paid by households, since they already had to pay energy taxes. In 1995, the tax on CO_2 was raised to 80 euros per ton, but in exchange for a cut in social insurance contributions. Such tax substitution was intended to reap a double dividend, i.e., to reduce both greenhouse gas emissions and tax distortions on the labor market. Germany, The Netherlands, Norway, Sweden, and the UK enacted similar tax reforms in the 1990s and early 2000s (see table 7.2). In 2009, the French government tried and eventually failed to introduce a carbon tax that was supposed to be fully compensated. The tax was rejected both by the Constitutional Council and by industrial lobbies.

In some cases, Pigovian taxes (see definition above) can be extremely effective, provided the tax rate is high enough. In 2002, for instance, Ireland introduced a heavy levy on plastic bags (0.15 euros per bag). By the end of the year, the consumption of these bags had fallen by 90% (OECD, 2007).[42] In 1991, a heavy tax on CO_2 and SO_2 emissions was introduced in Sweden. The subsequent reduction in emissions exceeded 50%. In Norway, the carbon emission tax introduced in 1991 led to a reduction in corresponding emissions by 21% the same year.

42. In the Irish case, retailers were obliged to fully pass the tax on to their customers.

Table 7.2
Green tax reforms in the 1990s and early 2000s

Country	Start year	Taxes raised on	Tax cut	Magnitude
Denmark	1994	Various[a] CO_2, SO_2	Personal income tax Social insurance contributions Capital income	Around 3% of GDP by 2002 (6% of total tax revenue)
Germany	1999	Petroleum products	Social insurance contributions	Around 1% of total tax revenue in 1999 and 1.8% in 2002
The Netherlands	1996	CO_2	Corporate income tax Personal income tax Social insurance contributions	0.3% of GDP in 1996, or around 0.5% of tax revenues in 1999
Norway	1999	CO_2, SO_2 Diesel oil	Personal income tax	0.2% of total tax revenue in 1999
Sweden	1990	Various* CO_2, SO_2	Personal income tax Energy taxes on agriculture Continuous education	2.4% of total tax revenue
UK	1996	Landfill	Social insurance contributions	Around 0.1% of total tax revenue in 1999
UK	2001	Energy (for industry)	Social insurance contributions	0.2% of total tax revenue in 2002

Note: [a]Gasoline, electricity, water, waste, and cars.
Source: CESifo DICE Report 3/2007, p. 46, from OECD (2007).

By construction, however, success with a Pigovian tax reduces the revenue that can be expected from this tax, so any "double dividend" is unlikely in practice. An emblematic example outside the environmental sphere is that of the *Tobin tax**, a small tax on capital flows inspired by Nobel Laureate James Tobin (1978)[43] that has been advocated by a number of NGOs to limit the scope for speculation and, at the same time, to raise revenues for less-developed countries. The success of this tax in reducing turnover on capital markets would, however, have defeated the second objective of raising revenues. In 2006, a number of countries, led by Brazil and France, decided to raise a new levy on airline tickets and to use the revenues from the tax to fight diseases in low-income countries. In this case, the objective was not to discourage people from flying, but rather to raise revenues on a relatively inelastic tax base. Again, there was no double dividend, just tax revenues.[44]

Despite their being relatively effective, environmental taxes face two difficulties related to their impact on redistribution and on competitiveness.

First, so-called "green" taxes are generally found to be regressive: A poorer household spends a larger share of its income on heating and transportation. Governments are thus tempted to provide poor households with targeted benefits that, of course, reduce the effectiveness of environmental taxation. Conversely, compensating poor households through raising means-tested benefits allows the government to correct the redistributive effect of the tax while preserving its effectiveness.

Second, by construction, green taxes raise the costs incurred by environment-intensive (generally energy-intensive) industries. To the extent that these industries compete worldwide, environmental taxes tend to reduce the competitiveness of domestic production; in turn, this reduces the global effectiveness of the tax since pollution is "imported" rather than produced domestically (see Copeland and Taylor, 2003). To circumvent these problems, governments often grant tax exemptions, but this amounts to giving up Pigovian taxation on the largest polluters. Conversely, governments can compensate firms through lump-sum taxation or through cuts in other taxes (especially social insurance contributions). Depending on how these compensatory transfers or tax cuts are designed, there may be a large redistribution effect across sectors. Finally, suggestions have been made that compensatory tariffs should be levied on imports from countries that do not

43. After an old suggestion by Keynes (1936): "The introduction of a substantial Government transfer tax on all transactions might prove the most serviceable reform available, with a view to mitigating the predominance of speculation over enterprise in the United States." John Maynard Keynes (1936), chapter 12, VI.

44. In the wake of the 2007–09 global financial crisis, a debate emerged on whether a new levy should be imposed on systemically important financial institutions, with diverging views on whether the *systemic levy** should be "Pigovian" and aim at reducing their propensity to leverage and take risk, whether it should be designed as an insurance premium imposed on too-big-to-fail institutions, or whether it should just aim at raising revenues to finance global public goods or national budgets.

comply with the *Kyoto protocol**,[45] either in cash or by forcing exporting companies to buy emission credits (*carbon inclusion mechanism**).

Environmental taxes, however, cannot be simply derived from optimal Pigovian tax considerations, for they are also influenced by:

- The possibility of reaching environmental objectives through alternative instruments: When the marginal cost of de-pollution is very uncertain, quantitative instruments (i.e., norms or emission markets) are a more effective way of controlling the volume of emissions. Additionally, in some cases quantitative norms involve smaller control costs than do Pigovian taxes.
- Political economy considerations, as taxes involve a number of contradictory parochial interests. Industrial lobbies resist taxation and organize themselves to propose voluntary contributions in order to rule out the alternative of regulation or taxation (see Wilson, 1980). The capacity of various countries to raise environmental taxes also depends on "objective" factors such as: The share of polluting industries in domestic output, social preferences, the intensity of foreign competition, geography, etc. For instance, the size of the US and its relatively low population density may explain why its citizens, confronted with large transportation needs, are attached to cheap energy and oppose energy taxes.

7.3.4 Tax cooperation

The debate on tax competition opposes those who praise its positive effect on government efficiency, and those who accuse it of distorting public choices and inducing inequality. The underlying paradigm behind the first argument is that of the *Leviathan government**, namely a partisan government moved by electoral objectives or dominated by an administration plagued by its own logic; the opponents, in contrast, believe in a *benevolent government** whose objectives coincide with social ones and are not taken hostage by the administration. These two polar visions coexist in Europe. Tabellini and Wyplosz (2004) provide a tentative synthesis:

> All this assumes that tax competition is undesirable. But is it? Not always and everywhere. If tax competition limits the tendency for governments to become overlarge, this may be welcome. International comparisons hardly suggest that the growth of the public sector in Europe is stunted by obstacles

45. Under the Kyoto protocol, a number of advanced economies have committed to reducing their emissions of greenhouse gases compared to their 1990 levels. They can use any instrument to reach these objectives. In 2005, the EU launched a tradable emission permit system so as to reduce the cost of complying with its commitments within the Kyoto protocol, compared to either using taxes or quantitative norms.

> to revenue. Moreover, heterogeneity of preferences remains an important reason to oppose centralization in tax matters.
>
> Guido Tabellini and Charles Wyplosz, 2004, pp. 26–27

Tax coordination within the EU is impeded by strong disagreements regarding the degree of desirable tax competition. Although nobody would go as far as proposing a harmonization of personal income taxation, coordination on corporate taxation gives rise to lively disputes between, for example, France and Germany, which tend to favor tax coordination, and the UK, Ireland, and Poland, which tend to oppose it. However, the decision-making process of the European Council requires unanimity for tax issues, which, in practice, blocks (or considerably slows down) any cooperation initiative on tax matters. Somewhat paradoxically, the only example of strong coordination in the EU concerns VAT, even though it mostly affects immobile tax bases (see box 7.17). This situation can be explained by the desire of EU member states to eliminate distortions on the "single market": VAT harmonization is viewed as a useful complement to the single market in goods and services, whereas capital tax harmonization is not yet fully viewed as a useful complement to the single capital market.

Still, the European Commission has been very active in promoting capital tax harmonization. A "tax package" was adopted in January 2003 that includes a "code of conduct" regarding detrimental practices on corporate taxation (for instance, tax rebates for foreign-owned companies)[46] and full exchange of information across member states on capital income, after a transitory period during which countries that still apply bank secrecy (Austria, Belgium, Luxembourg) have agreed to apply a withholding tax.

Simultaneously, some initiatives have been taken on corporate taxation. In 1990, the "mother–affiliate" directive tackled the double taxation of repatriated profits by a mother company from its subsidiaries. Member states are requested either to exempt repatriated profits, or to deduct taxes already paid by the affiliates from the mother's tax bill (partial credit system). The objective was to avoid discriminating against foreign subsidiaries (taxed twice) in relation to local firms (taxed only once).

In 2001, the European Commission proposed a two-step strategy to remove remaining corporate tax distortions in the EU: On the one hand, to suppress specific distortions (for example, by extending the scope of the mother-affiliate directive); on the other hand, to harmonize and consolidate the corporate tax base across member states through a *Common, Consolidated Corporate Tax Base** (CCCTB) system, i.e., through consolidation and apportionment of the tax base.

Such an apportionment system is used in Canada and the US. The European Commission has proposed to introduce it in the EU. According to

46. According to the Primarolo report (1999), there were 66 "detrimental practices" within the EU. An agreement was reached to dismantle these practices and to avoid creating new ones.

the CCCTB, each member state would be allocated a share of the single consolidated tax base of each multinational firm or, alternatively, of the single, EU-wide CIT base, according to some apportionment formula, based on physical capital, payrolls, turnover, or a combination of the three. It could then tax this base at its own statutory rate. Many details such as the scope of the consolidation, the definition of tax allowances, etc., need to be determined. Although a consensus is unlikely to emerge on this issue, such a system might also be partially introduced through the enhanced cooperation scheme.[47] However, such a CCCTB would likely strengthen rather than dampen tax competition due to higher transparency of tax rates, although the scope for tax optimization would be reduced. This raises the question of imposing a minimum tax rate, which is even less consensual than the CCCTB.

Box 7.17 VAT in the EU

VAT is the only tax subject to harmonization in the EU, as a complement to the single market. According to a 1977 directive, three different VAT rates are applicable in the EU: A standard rate (minimum 15%) and two reduced rates (minimum 5%). Some "super-reduced" rates (2% to 4%) can be seen as inheritances from the past to be progressively phased out, and some activities, such as financial services, are exempted from VAT. Reduced rates can be applied only to a limited list of subsistence items such as food or drugs. In 1999, the European Council extended the right to use reduced VAT rates for a strict list of labor-intensive services (small repairs, house renovation, house cleaning, domestic care, hairdressing) for an experimental period of three years, in order to boost job creation in these sectors. The experiment was then extended several times and in May 2009 the Council allowed these exceptions to remain permanent. Indeed, it is easy for member states to argue that, because those services are mostly immobile across EU countries, cutting VAT on them is not harmful to other member states, so the subsidiarity principle (see chapter 2) should apply to them.

Within the EU, VAT is raised according to the principle of destination, i.e., it is raised in the country where final consumption takes place, at that country's prevailing rate, except for "old" motor vehicles and distance sales, where the principle of origin applies. The destination principle raises a number of administrative problems (each producer faces 27 tax administrations) as well as fraud (carousel fraud, representing around 50 billion euros in 2005). Therefore, the destination principle was supposed to be a transitory device until a VAT based on the origin principle could be introduced. Today, however, the destination principle for VAT appears

47. See chapter 2.

as the last shield against tax competition across EU member states. Indeed, cutting VAT in one country does not provide any competitiveness advantage, and in the longer run, no more advantage than cutting social contributions (see section 7.3.1). When VAT is raised in the producers' country, however, competition on VAT will likely emerge. It could be seen as potentially limiting the tendency of governments to expand public expenditures or as rebalancing taxation away from consumption towards capital and considered as good news. However, in the absence of harmonization on other taxes, there is a risk that European governments would no longer be able to finance government services and to redistribute across individuals.

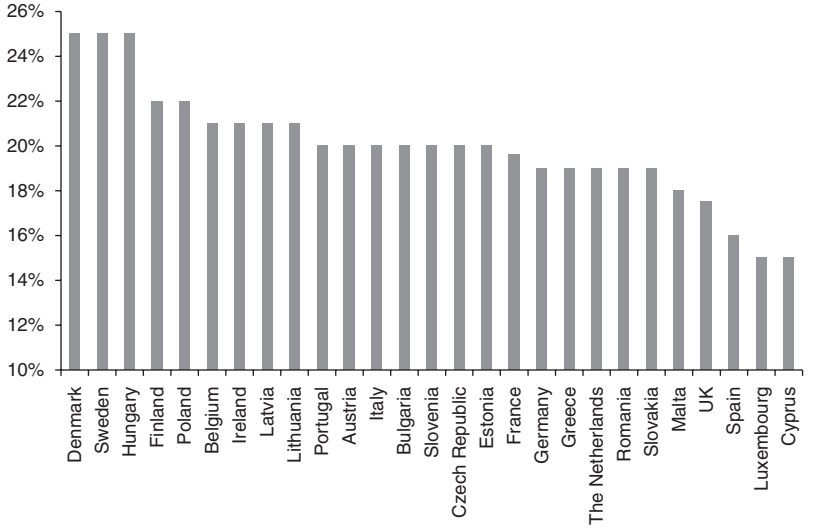


Figure B7.17.2 VAT standard rates in the EU at 1 January 2010.
Source: European Commission, DG Taxation and Custom Union, "VAT Rates Applied in the Member States of the European Union", taxud.d.1(2010) 118380.

Ultimately, tax competition in the EU raises the challenge of financing the European budget. As detailed in box 3.15 of chapter 3, the EU budget is financed through member states' contributions, transfers of VAT revenues, and import taxes. Even though the EU budget finances common policies, each member state is tempted to calculate its net contribution, i.e., the difference between what it gives to and what it gets from the EU budget. This blocks any discussion on the characteristics of the budget in terms of allocation and of redistribution (Tabellini, 2003). If, as the Sapir report recommended in 2004, the European budget had to develop towards more redistribution (between rich and poor regions, between expanding regions and regions in

conversion) and higher provision of government services (infrastructures, R&D), then it would be advisable to back the budget with a genuine European tax, paid by citizens or companies and not by states. In all logic, this tax should replace some existing taxes (because member states would reduce their direct contribution to the EU budget, or companies would be permitted to credit the European tax on the national tax), and it should rest on a mobile base within the EU such as corporate income (because it would allow internalizing tax externalities). Such ideas are regularly debated, but they encounter fierce political opposition from countries like the UK that oppose any loss of sovereignty on tax matters, and fear that the introduction of a European tax would be a prelude to an increase in the EU budget and to a widening of the remit of the EU.

References

Arulampalam, W., M.P. Devereux, and G. Maffini (2007), "The incidence of corporate income tax on wages," Oxford University Centre for Business Taxation working paper, 07/07.

Andersson, F., and R. Forslid (2003), "Tax Competition and Economic Geography," *Journal of Public Economic Theory*, 5, pp. 279–304.

Atkinson, A. (1970), "On the Measurement of Inequality," *Journal of Economic Theory*, 2, pp. 244–63.

Atkinson, A. (1977), "Optimal Taxation and the Direct versus Indirect Tax Controversy," *Canadian Journal of Economics*, 10, pp. 590–606.

Auerbach, A. (2005), "Who Bears the Corporate Tax? A Review of What We Know," NBER working paper, no. 11686, October.

Auerbach, A., and J. Hines (2002), "Taxation and Economic Efficiency," in Auerbach, A. and M. Feldstein (eds.), *Handbook of Public Economics*, vol. 3, North Holland, chap. 21.

Auerbach, A., M.P. Devereux, and H. Simpson (2007), "Taxing Corporate Income," CESifo working paper no. 2139, November.

Baldwin, R., R. Forslid, P. Martin, G. Ottaviano, and F. Robert-Nicoud (2003), *Economic Geography and Public Policy*, Princeton University Press.

Barnaud, L., and L. Ricroch (2005), "Les taux marginaux d'imposition: quelles évolutions depuis 1998?," *Diagnostic, prévisions et analyses économiques*, no. 63, March, Paris: French Ministry of Economy and Finance, available at http://www.minefi.gouv.fr/directions_services/dgtpe/TRESOR_ECO/tresorecouk.htm.

Bourguignon, F., and D. Bureau (1999), "L'architecture des prélèvements obligatoires en France: état des lieux et voies de réforme," *Rapport pour le Conseil d'Analyse Economique*, no. 17, Paris: La Documentation Française.

Bourguignon, F., and P.-A. Chiappori (1998), "Fiscalité et redistribution," *Revue Française d'Economie*, 13, pp. 1–64.

Brook, A.M., and W. Leibfritz (2005), "Slovakia's Introduction of a Flat Tax as Part of Wider Economic Reforms," OECD Economics Department Working Paper no. 448, Paris: Organisation for Economic Co-operation and Development.

Bucovetsky, S., and J.D. Wilson (1991), "Tax Competition with Two Tax Instruments," *Regional Science and Urban Economics*, 21, pp. 333–50.

Cahuc, P. (2002), "A quoi sert la prime pour l'emploi?," *Revue Française d'Economie*, 16, pp. 3–61.
Clausing, K.A. (2007), "Corporate Tax Revenues in OECD Countries," *International Tax and Public Finance*, 4, pp. 115–133.
Conseil National des Impôts (2003), "La fiscalité dérogatoire," Report to the President of the Republic available at www.ccomptes.fr.
Coase, R.H. (1937), "The Nature of the Firm," *Economica*, 4, pp. 386–405.
Copeland, B.R., and M.S. Taylor (2003), "Trade, Growth, and the Environment", *Journal of Economic Literature*, 42, pp. 1–71.
De Mooij, R.A., and S. Ederveen (2001), "Taxation and Foreign Direct Investment: A Synthesis of Empirical Research," *International Tax and Public Finance*, 10, pp. 673–93.
De Nardi, M., Liqian Ren, and C. Wei (2000), "Income Inequality and Redistribution in Five Countries," *Economic Perspectives*, 4, 2nd quarter, Chicago: Federal Reserve Bank of Chicago.
Devereux, M.P. (2006), "The Impact of Taxation on the Location of Capital, Firms and Profit: Survey of Empirical Evidence," Oxford University Centre for Business Taxation working paper 07/02.
Devereux, M.P. (2007), "Developments in the Taxation of Corporate Profit in the OECD since 1965: Rates, Bases and Revenues," Oxford University Centre for Business Taxation working paper 07/04.
Devereux, M.P., and R. Griffith (2002), "The Impact of Corporate Taxation on the Location of Capital: A Review," *Swedish Economic Policy Review*, 9, pp. 79–102.
Devereux, M.P., and P.B. Sørensen (2006), "The Corporate Income Tax: International Trends and Options for Fundamental Reform," *European Economy*, Economic Papers, 264, December.
Devereux, M.P., R. Griffith, and A. Klemm (2002), "Corporate Iincome Tax Reforms and International Tax Competition," *Economic Policy*, 17, pp. 449–95.
Emran, M.S., and J. Stiglitz (2005), "On Selective Indirect Tax Reform in Developing Countries, *Journal of Public Economics*, 89, pp. 599–623.
European Commission (2001), "Company taxation in the internal market," Staff Commission working paper, COM (2001) 582, Brussels: European Commission.
European Commission (2007), "Taxation Trends in the European Union," 2007 edition, Luxembourg: Eurostat.
European Commission (2009), "Taxation Trends in the European Union," 2009 edition, Luxembourg: Eurostat.
Feenstra, R.C. (2004), *Advanced International Trade*, Princeton University Press.
Feld, L.P., and E. Reulier (2008), "Strategic Tax Competition in Switzerland: Evidence from a Panel of the Swiss Cantons," *German Economic Review*, 10, pp. 91–114.
Feldstein, M. (1978), "The Welfare Cost of Capital Income Taxation," *Journal of Political Economy*, 86, pp. 29–51.
Feldstein, M. (2005), "Raise Taxes on Savings? Tell Joe It Ain't So!, *The Wall Street Journal*, 8 December.
Feldstein, M. (2008), "Effects of Taxes on Economic Behavior," *National Tax Journal*, 61, pp. 131–39.
Fjeldstad, O.-H., and L. Rakner (2003), "Taxation and Tax Reforms in Developing Countries: Illustrations from sub-Saharan Africa," CMI Report, R 2003: 6, Bergen: Chr. Michelsen Institute.
Friedman, M. (1962), *Capitalism and Freedom*, University of Chicago Press.

Godefroy, R. (2003), "Les taxes sur les cigarettes sont-elles régressives?," *Economie Publique*, no. 13, pp. 3–28.
Grossman G., and H. Helpman (1994), "Protection for Sale," *American Economic Review*, 84, pp. 833–50.
Graetz, M.J. (2007), "Tax Reform Unraveling," *Journal of Economic Perspectives*, 21, pp. 69–90.
Harberger, A. (1962), "The Incidence of the Corporate Income Tax," *Journal of Political Economy*, 70, pp. 215–40.
Hines, J.R. (1999a), "Three Sides of Harberger Triangle," *Journal of Economic Perspectives*, 13, pp. 167–88.
Hines, J.R. (1999b), "Lessons from Behavioral Responses to International Taxation," *National Tax Journal*, Vol. 2, pp. 305–23.
Hines, J.R. (2007), "Corporate Taxation and International Competition," in Auerbach, A.J., Hines, J., and Slemrod, J., (eds.), *Taxing Corporate Income in the 21st Century*, Cambridge University Press, pp. 268–95.
Hodge, S.A., S. Moody, and P.W. Warcholik (2005), "The Rising Cost of Complying with the Federal Income Tax," Tax Foundation, Special Report no. 138.
Holt, S. (2006), "The Earned Income Tax Credit at Age 30: What We Know," *Brookings Institution Research Brief*, February.
Hufbauer, G., and K. Elliott (1994), *Measuring the Cost of Protection in the United States*, Institute for international Economics.
Hume, D. (1752), "Of Taxes," in *Essays, Moral, Political, and Literary*, Cosimo Classics, 2007, pp. 349–54.
Ivanova, A., M. Keen, and A. Klemm (2005), "The Russian Flat Tax Reform," IMF working paper no. 05/16, Washington DC: International Monetary Fund.
Joumard, I. (2002), "Tax Systems in European Union countries," *OECD Economic Review*, 34, pp. 97–164, Paris: Organization for Economic Cooperation and Development.
Kahneman, D., and A. Tversky (2000), *Choices, Values and Frames*, Cambridge University Press.
Keen, M., and S. Smith (2006), "VAT Fraud and Evasion: What Do we Know, and What Can Be Done?," *National Tax Journal*, 59, pp. 861–87.
Keen, M., Y. Kim, and R. Varsano (2007), "The Flat Tax(es): Principles and Experience," *International Tax and Public Finance*, 15, pp. 712–51.
Keynes, J.M. (1936), *The General Theory of Employment, Interest and Money*, Harcourt, Brace & Co., chapt. 12,
Kirchgässer, G., and W. Pommerehne (1996), "Tax Harmonization and Tax Competition in the European Union: Lessons from Switzerland," *Journal of Public Economics*, 60, pp. 351–71.
Krogstrup, S. (2002), "What Do Theories of Tax Competition Predict of Capital Taxes in EU Countries? A Review of the Literature," HEI Working working paper no. 05/2002.
Laffer, A. (2004), "The Laffer Curve: Past, and Present Future," Backgrounder no. 1765, Washington DC: The Heritage Foundation, available at www.heritage.org.
Leiserson, G., and J. Rohaly (2006), "The Distribution of the 2001–2006 Tax Cuts: Updated Projections, November 2006," Tax Policy Center, Urban Institute and Brookings Institution, available at www.taxpolicycenter.org.
Lépine, J.-L., P.F. Gouiffès, and J. Carmona (1999), "Mission d'analyse comparative des administrations fiscales: rapport de synthèse," Paris: Inspection Générale des Finances, available on www.ccomptes.fr.

Lusardi, A., and O. Mitchell (2007), "Baby Boomer Retirement Security: The Roles of Planning, Financial Literacy, and Housing Wealth," *Journal of Monetary Economics*, 54, pp. 205–24.

Malinvaud, E. (1998), "Les cotisations sociales à la charge des employeurs: analyse économique," Report du Conseil d'Analyse Economique no. 33, Paris: La Documentation Française.

Marchese, C. (2007), "A Chinese Receipt for Curbing the Evasion of Commodity Taxes?," CESifo DICE report no. 3/2007.

Mathis, A. (2004), "VAT indicators," European Commission Taxation Papers no. 2/2004.

Mirrlees, J. (1971), "An Exploration of the Theory of Optimal Income Taxation," *Review of Economic studies*, 38, pp. 175–208.

Musgrave, R. (1997), "Reconsidering the Fiscal Role of Government," *American Economic Review*, 87, pp. 156–59.

OECD (2006), "Reforming Personal Income Tax," *Policy Brief*, March.

OECD (2007), "The Political Economy of Environmentally Related Taxes," February.

Olson, M. (1969), "The Principle of 'Fiscal Equivalence': The Division of Responsibilities among Different Levels of Government," *American Economic Review*, 59, pp. 479–87.

Overesch, M. (2005), "The Effective Tax Burden of Companies in Europe," CESifo DICE Report 4/2005, pp. 56–63.

Pigou, A.C. (1920), *The Economics of Welfare*, Macmillan & Co.

Piketty, Thomas (2001a), *Les Hauts Revenus en France au 20ème siècle: inégalités et redistribution, 1901–1998*, B. Grasset.

Piketty, Thomas (2001b), "Les inégalités dans le long terme," in *Inégalités économiques,* Rapport du Conseil d'Analyse économique, no. 33, Paris: La Documentation Française, pp. 137–204, available on www.cae.gouv.fr.

Piketty, T., and E. Saez (2007), "How Progressive is the US Federal Tax System? A Historical and International Perspective," *Journal of Economic Perspectives*, 21, pp. 3–24.

Primarolo, D. (1999), "Code of Conduct (Business Taxation)," Report to the ECOFIN Council, 23 November, available on the European Commission (Taxation and Customs Union) Web site, http://ec.europa.eu/taxation_customs/taxation/company_tax/harmful_tax_practices/.

Ramsey, F.P. (1927), "A Contribution to the Theory of Taxation," *Economic Journal*, 37, pp. 47–61.

Salanié, B. (1998), "Un exemple de taxation optimale," in F. Bourguignon (ed.), *Fiscalité et reduistribution*, Rapport du Conseil d'Analyse Economique no. 11, pp. 87–90, Paris: la Documentation Française.

Salanié, B. (2003), *The Economics of Taxation*, MIT Press.

Sandmo, A. (1985), "The Effects of Taxation on Savings and Risk Taking," in Auerbach, A. J., and Feldstein, M. (eds.), *Handbook of Public Economics*, vol. 1, North-Holland, pp. 265–311.

Sapir, A., P. Aghion, G. Bertola, M. Hellwig, J. Pisani-Ferry, D. Rosati, J. Viñals, and H. Wallace (2004), *An Agenda for Growing Europe*, report to the President of the European Commission, Oxford University Press.

Slemrod, J. (1990), "Optimal Taxation and Optimal Tax Systems," *Newspaper of Economic Perspectives*, 4, pp. 157–78.

Slemrod, J., and S. Yitzhaki (2002), "Tax Avoidance, Evasion, and Administration," in *Handbook of Public Economics*, chapter 22, pp. 1423–70.

Sterdyniak, H., M.H. Blonde, G. Cornilleau, J. Le Cacheux, and J. Le Dem (1991), *Vers une Fiscalité Européenne*, Economica.

Stern, N. (1987), "Optimal Taxation," in Eatwell, J., Milgate, M., and Newman, P. (eds.), *The New Palgrave. A Dictionary of Economics*, vol. 3 (K–P), Macmillan.

Stigler, G. (1946), "The Economics of Minimum Wage Legislation," *American Economic Review*, 36: 358–65.

Tabellini, G. (2003), "Principle of Policymaking in the European Union," *CESifo Economic Studies*, 49, pp. 75–102.

Tabellini, G., and C. Wyplosz (2004), *Réforme structurelle et coordination en Europe*, Rapport du Conseil d'Analyse Economique no. 51, Paris: La Documentation Française.

Thaler, R., and C. Sunstein (2003), "Libertarian Paternalism," *American Economic Review*, 93, Papers and Proceedings of the 115th Annual Meeting of the American Economic Association, Washington, DC, pp. 175–79.

Tiebout, Ch. (1956), "A pure theory of local expenditures," *Journal of Political Economy*, 64, pp. 416–24.

Tobin, J. (1978), "A Proposal for International Monetary Reform," *Eastern Economic Journal*, 3, pp. 153–59.

Wanniski, J. (2005), "Sketching the Laffer Curve," *Yorktown Patriot*, 14 June.

Wilson, J. (1980), *The Politics of Regulation*, Basic Books.

Zodrow, G., and P. Mieszkowski (1986), "Pigou, Tiebout, Property Taxation, and the Underprovision of Local Public Goods," *Journal of Urban Economics*, 19, pp. 356–70.

8

Economic Policy and the 2007–09 Crisis

8.1 What Went Wrong?
- 8.1.1 A brief account of the crisis
- 8.1.2 Three questions on the crisis
- 8.1.3 A taxonomy of crisis roots
- 8.1.4 Micro roots
- 8.1.5 Macro roots
- 8.1.6 The "Black Swan" syndrome
- 8.1.7 Lessons

8.2 Extraordinary Times
- 8.2.1 Economic policy without the usual compass
- 8.2.2 The aftermath
- 8.2.3 Lessons

8.3 In Search of a New Regime
- 8.3.1 The financial system
- 8.3.2 The macroeconomic policy regime
- 8.3.3 Conclusion

References

The previous chapters of this book were mostly written while the global economy was growing at a rapid, stable, and noninflationary pace. Whether this "great moderation" was a result of prudent, predictable macroeconomic policies, or merely of luck, was a matter of debate among academics.[1] Some form of consensus had however emerged, which resulted in a set of policies that were deemed favorable to growth and stability (the "augmented Washington consensus," see chapter 6) and which emphasized the benefits of rule-based policies and the need to eschew discretionary activism. It was also agreed, and enshrined in international agreements, that any significant government assistance to private firms operating on competitive markets was to be regarded with considerable suspicion.

1. See especially Romer (1999) and Blanchard and Simon (2001).

The financial crisis that started in the summer of 2007 and moved into a sharp, global economic crisis in the autumn of 2008—and which we shall call in this chapter "the crisis"—suddenly led policymakers to break with the prevailing consensus. Not only did governments and central banks embark upon discretionary monetary and fiscal stimulus, but they also intervened heavily by bailing out banks and by assisting nonfinancial industries (especially the car industry).

The main reason why policymakers made this choice was probably that the memory of the Great Depression of the 1930s had not been lost. Even before it became clear that the fall in stock prices, output, and international trade was initially as fast as during the Great Depression, if not faster (Eichengreen and O'Rourke, 2009), policymakers decided to make full use of monetary and fiscal instruments to tackle the crisis. After the US investment bank Lehman Brothers went bankrupt in September 2008—with dire consequences for market conditions—they also put the free-market ethos aside and embarked on wholesale bank support.

As a result the crisis gave rise to what the heads of state of the G20 called "the largest and most coordinated fiscal and monetary stimulus ever undertaken."[2]

Financial crises are not exceptional events (see the historical record reported by Reinhart and Rogoff, 2009a,b) but truly global crises are. This one immediately triggered two debates.

The first debate has been about the causes of, and the responses to the crisis. It started early but is unlikely to be settled soon. On-the-spot analyses are often partial and overly influenced by particular aspects of the chain of events. It took decades to clarify why the Great Depression occurred: It was only in 1963, with the publication of Milton Friedman's seminal book with Anna Schwartz, *A Monetary History of the United States*, that the responsibility of monetary policy was highlighted, and it was in the 1980s—half a century after the fact—that Ben Bernanke brought new light to the debate with his research on the role of the credit channel. However, as in the 1930s, action had to be taken and was taken without delay, on the basis of the available evidence and the immediate reading of the factors behind the crisis. This amounted to curing the consequences, not the causes, of the crisis.

The second debate was best captured by the Queen of England when she famously asked during a visit to the London School of Economics "why did no one see it coming." It has mostly developed among economists and has centered on the profession's potential responsibility for not having pointed out adequately that financial developments in the 2000s involved significant risks.

This chapter focuses on the first debate and hints at the second one. It does not attempt to provide a unified, empirically grounded analysis leading

2. According to the declaration of the September 2009 G20 summit of Pittsburgh (all G20 declarations are available at the G20 Information Center of University of Toronto, www.g20.utoronto.ca).

to unambiguous prescriptions. More modestly, we outline what we think we have learned thus far, what are the policy issues raised by the response to the crisis, and which are the longer-term priorities for reform. In addition, we show how the toolbox provided in chapters 1–7 of this book can be used to understand the crisis.

8.1 What Went Wrong?

8.1.1 A brief account of the crisis

The crisis started in a small and relatively obscure corner of the US mortgage credit market—the now world-famous subprime market. *Subprime mortgages** are financial products that aim to give access to home ownership to poorer and therefore less-creditworthy households. These high-yield mortgages are riskier, and contracts were designed so as to mitigate this risk thanks to rising house prices: Low-income borrowers could finance and refinance their homes by collateralizing them. This worked as long as house prices were rising, but in 2006 default rates started to ratchet up in response to the decline in house prices.

This would have remained the lenders' problem, had subprime credits not been securitized, i.e., transformed into marketable bonds (see box 8.1 for a description of securitization). Furthermore, they had also been pooled with other, higher-quality mortgage-based securities to form structured assets that were therefore riskier, and had a higher return than standard fixed-income instruments.[3] These securities were composed of tranches of declining quality and increasing risk, from the senior tranches that were rated AAA to the equity tranches. Only the latter (and possibly the intermediate ones) were supposed to be affected by subprime default, but as default rates exceeded what had been considered probable, the senior tranches were affected too. Asset-backed securities were further assembled, packaged, and then sliced again into tranches to form increasingly complex and opaque products. Packaging of this sort was commonplace, which explains why defaults on the subprime segments affected the whole range of asset-backed securities. Complex financial products previously considered safe became increasingly difficult to value. Asset holders became unable to value the "toxic" products they held on their balance sheet, let alone sell them.

Banks in the US and in Europe not only had invested in these assets, which had turned out to be riskier than first thought. They had also done so by issuing debt rather than by investing their own capital, largely through legally distinct subsidiaries (the so-called *conduits** and *special investment vehicles** or *SIVs*) that used the income stream from their assets to service their debt. Being squeezed between losses on asset-backed securities on the one hand,

3. The process of securitization and the main structured assets mentioned above are explained in box 8.1.

and (as their losses started being known) an increasing difficulty to roll-over their debts on the other hand, these so-called "shadow banks" (see below) had no choice but to draw on the credit lines they had with their parent banks. The latter then had either to extend credit to their subsidiaries or to repatriate them onto their balance sheets, and to seek a way to refinance them. However, in 2007 this became increasingly difficult because of rising mutual suspicion on the interbank market.

In August 2007, the usually highly liquid interbank market suddenly froze. Europe was affected as much as the US, because a large part of the so-called toxic assets had been bought by European banks. Central banks instantly stepped in and started to play their role as lender of last resort, providing liquidity directly to financial institutions (against collateral, see the practicalities in chapter 4) in order to help them face debt repayment schedules. However, liquidity provision was not enough to restore confidence, because markets participants suspected that some counterparties were potentially bankrupt and were unwilling to lend to them.

Losses were meanwhile compounded as banks started to sell assets for which there was still a market—frequently stocks—to reap liquidity and comply with capital ratios. The resulting fall in asset prices in turn further damaged the banks' balance sheets, as these are based on the market values of assets (this is known as *mark-to-market accounting**) and the fall in asset prices forced banks to sell further assets. Furthermore, many complex assets they had purchased were no longer being traded and published accounts therefore did not provide accurate information on the true extent of the damage. As a result, some banks were proved, or suspected, to be insolvent, which exacerbated mistrust in the interbank market. The demise of Northern Rock, a UK building society which asked for liquidity support from the Bank of England in September 2007 and was subsequently taken into state ownership, illustrated the consequences of the liquidity crisis.

The panic reached a climax in September–October 2008 in the wake of incoherent responses by US authorities—investment bank Bear Sterns and insurer AIG were bailed out, but Lehman Brothers, another investment bank, had to default—and the bail-out of Fortis and Dexia, two major European banks with complex cross-border operations. There was a massive loss of confidence. Everybody hoarded liquidity and central banks had to cut interest rates to zero and engage in a near-total substitution of the interbank market.

At this stage contagion to the real economy amplified: As economic confidence plummeted, companies started to postpone investments and reduce inventories; the fall in equity prices and the freeze of corporate bond markets reduced the ability of large companies to finance their investments; and households responded to the shock with an increase in precautionary savings. Banks also became reluctant to lend to nonfinancial customers, since this would have raised their exposure to risk whereas they wanted to reduce it;

but there is little hard evidence on the existence of a generalized credit crunch, especially as governments soon stepped in to help banks continue lending.

Much more than at the time of the Great Depression, globalization led to a quasi-instantaneous international transmission of the shock. Starting in autumn 2008, banks reduced their exposure to emerging and developing markets, through rationing credit to their local subsidiaries (especially in Central and Eastern Europe). More generally, there was a "sudden stop" of capital outflows from the US and Europe. This was a crucial channel of crisis contagion to those emerging economies that relied on external financing. The other main channel was international trade: Cuts in investment and consumption plans, together with the reduction in inventories, and the drying-up of trade finance, dramatically reduced world trade. This especially affected East Asian countries whose growth models were based on export demand from the US and Europe, rather than on domestic or regional demand. The fall in previously inflated commodity prices also affected several emerging and developing countries. More generally, contracting demand in developed economies dragged the whole world into a recession, including low-income countries.

Governments at this stage responded to the crisis with full force. The US and Europe put in place bank rescue and guarantee plans amounting to about one-quarter of GDP. In an attempt to prevent further collapses they bailed out or nationalized insolvent banks, recapitalized the weak ones, and provided credit guarantees to all. Major budgetary stimulus plans soon followed, while several central banks, having cut interest rates to zero or near-zero levels, engaged in nonconventional easing measures. The Federal Reserve extended swap lines to a series of central banks around the world to help them counter the shortage of dollar liquidity. The IMF, the World Bank, regional development banks, and other donor institutions were also mobilized to counter capital outflows from emerging economies, finance international trade, and help developing economies engineer counter-cyclical policies. All this was not enough to prevent a world recession, but after a sharp fall of production in winter 2008–09 stabilization occurred in spring 2009.

These various steps of crisis contagion are summarized in table 8.1.

8.1.2 Three questions on the crisis

These developments raise three major questions: Why did the crisis occur? Why did it engulf the entire financial system? Why have its economic consequences been so severe?

The *third question* is the easiest to answer. Financial crises affect the real economy through credit supply constraints (this is the *credit channel* introduced in chapter 4), wealth effects (the drop in asset prices reduces

Table 8.1
Main stages in financial crisis development

Date	Events	Policy responses
2006–summer 2007	Localized credit concerns in the US	
	• Rising defaults in riskier housing mortgages • Falling prices of lower credit tiers of some credit securities	
Summer–autumn 2007	Initial cracks in confidence and liquidity strains	
	• Interbank rates rise sharply. Funding of asset-backed securities dries up • Failure of two large hedge funds • Run on UK bank Northern Rock	• Central banks extend liquidity to banks through exceptional tenders • Rescue of Northern Rock
Autumn 2007–early summer 2008	Accumulation of losses and continuation of liquidity strains	
	• Severe mark-to-market losses in trading books • Collapse of commercial paper market • Structured Investment Vehicles (SIVs) brought back on bank balance sheets • Worries about liquidity of major financial institutions	• Continued liquidity support by central banks • US government bails out investment bank Bear Stearns and sells it to JP Morgan
Summer 2008	Intensification of losses and liquidity strains	
	• Mark-to-market losses and liquidity strains escalate • US agencies Fannie Mae and Freddy Mac insolvent • Funding problems of UK mortgage banks intensify	• Fannie Mae and Freddy Mac de facto nationalized in early September

September 2008	Massive loss of confidence	
	• Bankruptcy of US investment bank Lehman Brothers • Loss of confidence that major institutions are too big to fail • Bankruptcy of Washington Mutual in the US, Bradford and Bingley in the UK, Icelandic banks • Almost total seizing up of interbank money markets and short-term funding markets	• US government refuses to bail out investment bank Lehman Brothers. Lehman Brothers files for bankruptcy protection. • US government bail-out of insurer AIG • Rescue of European banks Dexia and Fortis
October 2008		
		• Widening of collateral range and wholesale liquidity support by central banks • Governments assist banks through capital injections and funding guarantees • Explicit commitment that systemic banks will not be allowed to fail • Central banks' refinancing rates brought to zero or close to zero
Autumn 2008–spring 2009	Crisis transmitted to real economy	
	• Sharp decline in industrial production and GDP • Series of financial crises in emerging Europe as capital flows suddenly stop • Collapse of world trade • Slow normalization of interbank markets	• Central banks turn to unconventional policies • Large-scale government stimulus • International coordination of crisis responses • International swap agreements • IMF-led assistance programs

Source: Adapted and updated from Financial Services Authority (2009).

household wealth and diminishes consumption, while companies incur losses on their balance sheets and reduce investment accordingly), and, last but not least, confidence effects. A robust stylized fact emerging from a series of financial crises in recent decades is that they result in sharp and more or less prolonged drops in output (Cerra and Saxena, 2008; Reinhart and Rogoff, 2009a,b).

In 2008–09 international dimensions added to the shock and compounded its effects, resulting in the first global crisis since the end of the previous wave of globalization, in the 1930s. Although there are questions for research on the relative importance of the transmission channels and the magnitude of the corresponding effects, once the financial system had reached near-paralysis a sharp drop in global output had to be expected and initial hopes for a *decoupling** of emerging economies were soon rebuffed.[4]

The *first* and *second* questions—why a financial crisis, and why so widespread—are much more challenging. Part of the explanation can be found in financial conditions that prevailed in 2007, especially a high appetite for yield and a pervasive mispricing of risk, which had led many private financial agents to enter on a massive scale into debt-financed (or *leveraged*, see below) investments in risky assets. Once liquidity dried up and risk was re-priced, the same firms whose aim had been to maximize return through leverage entered into a precipitate and disorderly process of *deleveraging*.[5] These are standard developments in a financial crisis. However while it is easy to understand why investors exposed to the subprime credit risk were hurt, it is more difficult to find out why the entire financial system was affected. Part of the explanation here has to do with the complexity and connectedness of the global financial architecture: The system *looked* able to absorb and diffuse shocks, and it had performed very well when facing a sectoral shock on the occasion of the "dotcom" crash, but in 2007–08 it turned out that it amplified and reverberated, rather than diffused the shock arising from the subprime crisis. Part also resulted from the sale by banks of their remaining liquid assets, namely stocks. These "fire sales" transferred the crisis to the stock market and thereby reduced the value of the remaining stock on the banks' balance sheet.

To blame excess leverage in the financial sector or benign neglect from policy authorities as a key causes of the crisis is however unsatisfactory, since the deeper reasons for such behavior still need to be understood. Any serious discussion on the policy responses indeed has to start from an analysis of the root causes of the crisis.

4. Much hinges of course on what "decoupling" is supposed to mean. Clearly, the crisis has demonstrated that emerging markets were importantly affected by the implications of the shock originating in the US. However, it became apparent in the first half of 2009 that big emerging countries would come out of the slump earlier and faster than the US or Europe.

5. The implications of leverage for return and risk are explained in section 8.1.4c.

8.1.3 A taxonomy of crisis roots

There are three, nonmutually exclusive approaches to the roots of the crisis (figure 8.1):

- A first strand of analysis emphasizes the *microeconomic roots of financial imprudence*. According to this approach excessive risk-taking and *leveraging* (i.e., debt-financed financial investment) on the part of financial players were rooted in inadequate incentives that in turn can be ascribed either to insufficient or, on the contrary, to inappropriate regulation. This approach points to regulatory reform as the main response. There are, however, diverse views on what the regulatory agenda should be: Some advocate a mere increase of capital ratios, while others envisage much wider-ranging reforms of the structures of the financial system. The debate also takes on a moral dimension as greed is regarded by public opinion as having been at the heart of financial excesses.
- A second approach claims that *the macroeconomic environment contributed to excessive leveraging and risk-taking*. Two main factors contributed to such a lax environment. First, the US and global monetary policy stances have been criticized as excessively expansionary, which favored extensive leverage and the mispricing of financial and real-estate assets. Second, the flow of foreign savings into the US (which had global current-account imbalances as its counterpart) resulted in a low level of *long-term* interest rates and in a surge in the demand for (seemingly) safe dollar-denominated assets. For the supporters of this view, the underlying macro factors need to be addressed if future crises are to be avoided.
- Finally, a third view is inspired by engineering and ecology. It posits that the problem did not lie so much with either specific micro deficiencies or macro factors, but rather with the *resilience of the financial system as a whole*. Instead of putting emphasis on fundamental causes, it sees the financial turmoil as a very low-probability event (a "black swan") in which a shock of limited magnitude set in motion a chain reaction that eventually resulted in a near-collapse. The policy implication is that the emphasis should be put on strengthening the robustness of the financial system *as a whole*.

8.1.4 Micro roots

By far the most popular explanation of the crisis was the irresponsibility and "reckless greed and risk-taking," as expressed by President-elect Obama in January 2009. Popular representations combine imprudence, voracity, felony, and corruption to depict what could be called a series of behaviors à la Bernard Madoff. However, unchecked greed was already pervasive in

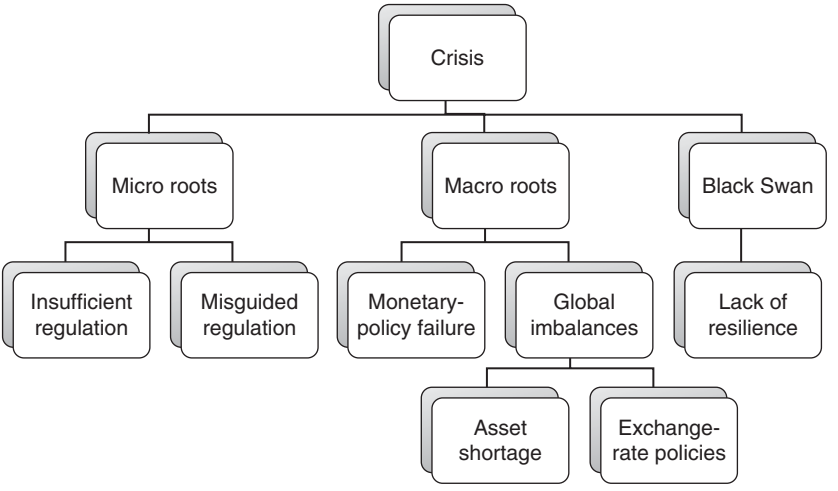


Figure 8.1 A simple taxonomy of crisis theories.

the 1980s and the 1990s while the financial system and the global economy prospered. Neither junk bonds nor the Enron fraud triggered a world crisis.

Scholars of economic policy must avoid a repetition of the error made at the time of the Asian crisis by those who blamed "crony capitalism" without questioning why cronyism, which had been there all along East Asia's path to prosperity, had suddenly become a problem.[6] And there is a thin line between "reckless greed" and self-interest, which economists have considered as the engine of decentralized economies since Adam Smith's famous remark that: "It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love." Sixty years earlier, Bernard de Mandeville had even argued that private vices were at the root of prosperity.[7]

A more interesting question is what led private-market participants to undervalue or misprice financial risk both on the sell-side and on the buy-side, and why an already burdensome public regulation and supervision apparatus did not tackle the problem. Since the crisis broke out, major deficiencies in what had become standard financial practice have been highlighted by observers. Many are important, raise puzzling questions, and call for significant regulatory reform. Four stand out: Compensation practices, securitization, leverage, and market valuation.

6. Paul Krugman (1998a) was an early advocate of the "cronyism" interpretation, in support of a widely shared view within the International Monetary Fund, before changing his mind about the causes of the Asian crisis (Krugman, 1998b).
7. Smith (1776, 1977); de Mandeville (1714).

a) Compensation practices

The traders' hefty bonuses have been widely resented by outraged citizens and they have become in some countries the symbol of the pre-crisis excesses. Beyond legitimate distributional concerns, the compensation structure impacts on incentives for risk-taking. In order to attract and retain talent, firms in the early 2000s routinely rewarded executives and traders on the basis of short-term performance. Executives generally received equity incentives in the form of options and shares without cash-out restrictions, and traders received bonuses tied to their expected performance (they were not adjusted *ex post* if the expected performance did not materialize). Also, a standard practice in banking was to reward executives with shares or options in a bank's parent holding company. Because the limited liability of shareholders restricted their potential losses to the value of their capital, managers had a strong interest in taking on leverage in order to maximize expected gains.[8] For all, compensation structures acted as a powerful incentive to take risk.

This issue is, at its core, one of corporate governance. Setting compensation is the role of a company board's compensation committee, which is expected to act in the interest of the holders of capital. However, as demonstrated in the crisis, the failure of a large bank or financial company involves systemic risk, which in turn compels public authorities to intervene to prevent it. This without doubt entails moral hazard and results in distorted incentives.

b) Securitization

Most economists consider that financial innovation is favorable to long-term growth (chapter 6) and that securitization is a case in point. The *packaging* of a series of loan portfolios into a single product and the *tranching* of this product into securities of different qualities of risk can be regarded as positive innovation: The former because it reduces dependence on specific portfolio risk and the latter because it allows investors to diversify and choose the desired combination of risk and return.[9] The basics of securitization are presented in box 8.1.

Box 8.1 A Primer on Securitization

*Securitization**, the technique through which bank loans are transformed into marketable securities, was invented in the 1970s when US-government-sponsored agencies like Fannie Mae (the Federal National Mortgage Association) started securitizing residential mortgages. Previously, banks held loans until they matured or were paid off (the so-called *originate and hold** model). But after World War II, depository

8. See the research by Lucian Bebchuk and colleagues, for example Bebchuk (2009).
9. For a discussion see Hellwig (2008).

institutions simply could not keep pace with the rising demand for housing credit and sought ways of increasing the sources of mortgage funding. To attract investors, an investment vehicle was developed that isolated mortgage pools, sorted them by order of credit quality and sold them as tranches, allowing banks to reduce their exposure to credit risk and thereby to increase their volume of credit. This is the so-called *originate-and-distribute** model.

Securitization implies the pooling of a large number of claims (such as mortgages, loans, bonds, trade, and credit-card receivables, etc.) and their use as collateral to issue a prioritized capital structure of claims (the *tranches**).[10] This process results in a series of rated securities. The highest tranches are senior to the lower ones, so that they can achieve a good risk rating even though the underlying collateral includes high-risk mortgages. The lower tranches are high-yield ones to compensate for the higher risk.

The best known such *asset-backed securities* (ABSs)* include *mortgage-based securities** (MBSs, collateralized by the service of mortgage loans), *collateralized mortgage obligations* (CMOs, emanating from a further securitization of MBSs), and the now-famous *collateralized debt obligations** (CDOs, resulting from the securitization of various ABS). Often, various credit-enhancement mechanisms are added to these products, such as *credit default swaps* (CDSs). In addition, CDOs were in turn sliced into tranches and sold to vehicles themselves financed by debt—thereby forming what was known as the "*CDO square*" or CDO^2.

Securitization was enormously successful: In the US, the amount outstanding of corresponding *asset-backed securities* reached 2.5 trillion dollars in 2007 (almost 20% of US GDP) and gave rise to further developments as simple securitized credits were restructured and repackaged into more complex ones.

However, sophisticated securitization had two consequences. First, it resulted in a major increase in the complexity of financial products that made risk difficult to assess. The first generation of structured products such as CDOs was admittedly rather simple, since their purpose was only to sort a bundle of loans into a series of tranches of increasing risk and expected return so as to match investors' different preferences for risk and return. However, even sophisticated investors had difficulty assessing and therefore monitoring the risk embedded in more complex products such as CDO^2, for two reasons: Details on the underlying risks were often not available, and even when they were, the value of the CDO was a complex, highly nonlinear function of the distribution of the underlying risks. Scrutiny of risk was widely outsourced to credit-rating agencies and more often than not replaced by a

10. For a full description and discussion, see Coval et al. (2009).

blind and ultimately lethal faith in the robustness of market mechanisms. As Buiter (2007) has noted, risk transferred through securitization ended up with the investors most *willing* to hold it, but not necessarily those most *able* to bear it.

The second consequence of securitization was that the credit originators—the lending institutions at the origin of the mortgage—had weak incentives to assess the credit risk. To the extent they were able to package and sell an entire credit portfolio, their incentive was limited to making sure that credit quality *as assessed at the time of the sale* matched the standards required by regulators and credit-rating agencies to qualify for a given risk category. The originate-and-distribute model of credit therefore involved moral hazard. Unsurprisingly, over the period following the sale of their loan, loans sold in the secondary market underperformed similar bank loans by a significant margin on a risk-adjusted basis (Berndt and Gupta, 2009). Securitization therefore contributed not only to disseminating risk, but also to creating new risk.

Additionally, the pooling of various loans in a single product was an efficient way to diversify individual risks but it did not allow diversification of the macroeconomic risk related to the housing bubble. When house prices started falling, a large number of borrowers were *simultaneously* unable to repay their debt. This rise in the correlation of individual default rates was not correctly taken into account in the models used by securitizers. The CDO tranches rated "AAA," the highest possible score, although they were deemed diversified enough to be robust, became vulnerable, triggering a loss of confidence and a contagion effect.[11]

c) Leverage

*Leverage** is a very old technique that makes it possible to increase the return on capital by incurring debt. Suppose an investor invests his or her capital K in a (risky) asset of expected yield r. The return s/he can expect to earn per unit of capital is then simply r. But if instead s/he borrows D at rate i and invests $A = K + D$ in the risky asset, s/he can expect to earn:

$$\rho = r + (r - i)l \tag{8.1}$$

where $l = D/K$ is the *leverage ratio**. When $i < r$, leverage thus appears as a simple way to increase return. Things are different when r turns out to be less than the cost of borrowing. And, worse, if the investor actually incurs a loss of z% on its investment, this implies a capital loss of z% without

11. See Coval et al. (2009) for a discussion of the role of correlations. The authors notably show that the high credit rating of many securities pointed to rating agencies being extraordinarily confident about their ability to measure the underlying default risks and default correlations. Small errors in evaluating the risk of underlying securities can however translate into substantial variation in the default risk of the final structured product.

leverage but of $(1 + l)z\%$ with leverage (and a total negative rate of return of $-[z + (z + i)l]$. The loss can exceed the investor's capital, which means s/he is unable to repay the debt and is therefore bankrupt.

Applied to banks, this simple mechanism has important consequences (Adrian and Shin, 2008). Even in the absence of a true bankruptcy, the very fact that a bank's assets have lost value implies a sudden rise in the leverage ratio, which is likely to lead the bank to sell off assets or restrict credit in order to deleverage. Suppose, for example, that initially $A = 100$, $D = 90$, and $K = 10$ (implying $l = 9$). Then a 5% decline in the value of A implies a 50% decline in the value of K, and thus a doubling of the leverage ratio. Bringing it back to its previous value of 9 implies a considerable shrinking of the balance sheet.

What this elementary calculation illustrates is the simple fact that leverage increases the expected return on capital but has two consequences: First, it also increases the risk of bankruptcy; second, it leads banks to respond *procyclically* to fluctuations in the value of their assets, thereby amplifying financial and economic fluctuations (box 8.2).

Box 8.2 Leverage and Procyclicality

Tobias Adrian and Hyun Song Shin (2008) have used micro-data to demonstrate the procyclicality of leverage in financial firms. Financial intermediaries adjust their balance sheets actively to changes in their net worth. Adrian and Shin first observe that for a passive investor, the relationship between the value of assets A and the leverage ratio l is downward-sloping: leverage falls when the value of total assets rises. This is simply because if debt D is held constant, l and A are negatively related:

$$l = \frac{D}{K} = \frac{D}{A - D} = \frac{1}{A/D - 1} \qquad \text{(B8.2.1)}$$

Data indicate that households follow this type of behavior as the relationship between asset growth and leverage growth is negative (figure B8.2.1).

This downward-sloping relationship gets lost for nonfinancial corporations. For commercial banks it becomes vertical at a zero-leverage growth intercept (figure B8.2.2): Commercial banks thus tend to keep leverage constant. This implies that debt is likely to be procyclical: Holding l constant means that the growth rate of debt D is the same as that of assets A.

The relationship is even reversed and turns positive for securities brokers and dealers, a statistical category that included the investment banks (figure B8.2.3), indicating strong leverage procyclicality: The higher the growth of total assets, the faster the growth of debt and of the

leverage ratio *l*. In other words, investment banks tended to accelerate borrowing when market conditions were improving. This is what led Lehman Brothers to excessive leveraged exposure to risky assets.

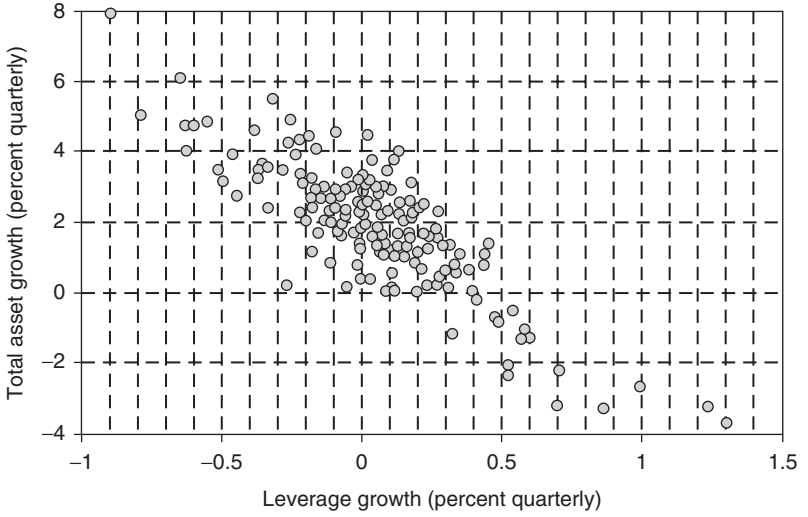


Figure B8.2.1 Relationship between asset growth and leverage growth, US households, 1963–2006.
Source: Adrian and Shin (2008, figure 2.2).

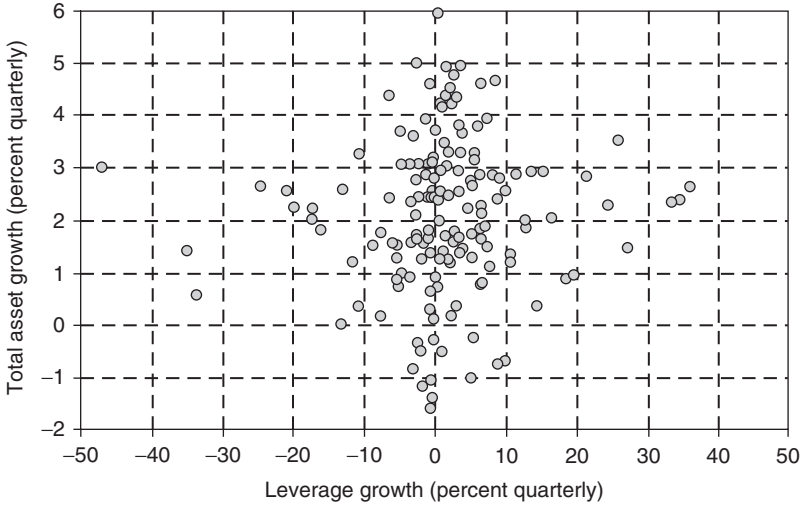


Figure B8.2.2 Relationship between asset growth and leverage growth, US commercial banks, 1963–2006.
Source: Adrian and Shin (2008, figure 2.4).

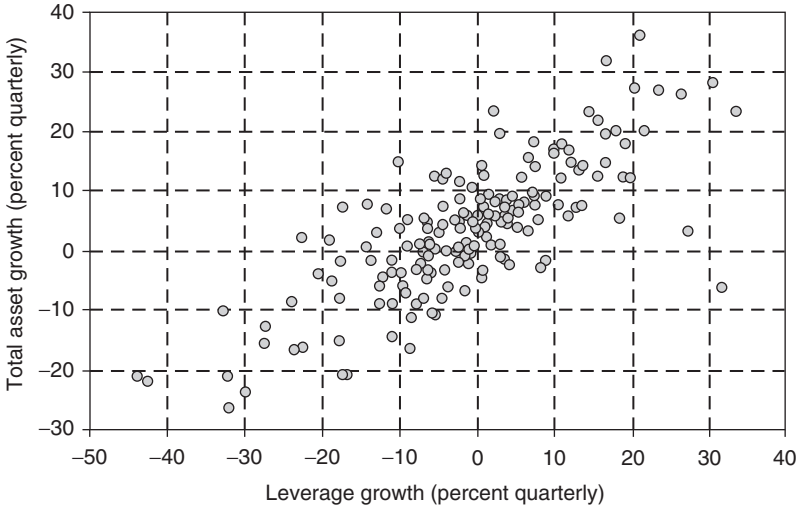


Figure B8.2.3 Relationship between asset growth and leverage growth, US securities brokers and dealers, 1963–2006.
Source: Adrian and Shin (2008, figure 2.5).

The procyclicality of leverage in turn results from the counter-cyclical behavior of measured risk (low during booms and high during busts). Adrian and Shin (2008) conjecture that banks maintain a stock of capital K proportional to total *value-at-risk* ($K = \lambda \times VaR$)—see chapter 2 for its definition. Using the same notation as previously, the leverage ratio l can be written as:

$$l = \left(\frac{A - K}{K}\right) = \frac{1}{\lambda}\frac{A}{VaR} - 1 \qquad \text{(B8.2.2)}$$

The leverage ratio l is therefore negatively related to unit Value-at-Risk, VaR/A. Adrian and Shin's data confirm the counter-cyclicality of unit Value-at-Risk, which implies the procyclicality of leverage. The interpretation is the following: When asset prices increase, financial intermediaries' balance sheets tend to get stronger, creating an incipient situation of surplus capital. The incentive is for intermediaries to find ways to employ this surplus capital through an expansion of balance sheets and an increase in leverage.

Given that large European banks in 2007 had leverage ratios between 20 in the UK and 35 in Switzerland,[12] these mechanisms played a major role in the transmission of the crisis from asset prices to bank behavior.

12. See Panetta and Angelini (2009).

As bank failures may involve massive externalities, leverage has to be regulated in order to limit excessive risk-taking. Thus an important question is why existing regulation failed. Part of the answer is to be found in the role of the *shadow banking** system (Adrian and Shin, 2009). By mid-2007, just before the crisis erupted, market-based assets amounted to more than 16 trillion US dollars, while bank-based assets were less than 13 trillion. Against this background, existing banking regulation proved insufficient. It mainly rested on two instruments: Mandatory capital adequacy ratios and leverage ratios.

Mandatory capital-adequacy ratios limit the size of a bank's risk-weighted asset portfolio relative to its capital. They are at the heart of the Basel I and Basel II accords (box 8.3) which by 2008 had been implemented in most OECD countries and many East Asian countries, but not in the US. Capital adequacy ratios proved to be both too low and unevenly enforced:

- In the US, neither hedge funds nor investment banks had to comply with capital-adequacy ratios (only bank holding companies had to), whatever the potential (or, in the case of Lehman, actual) repercussions of their bankruptcy;
- Most importantly, the sophisticated capital-adequacy ratios put in place after 2006 under the *Basel II accord** (see box 8.3) to replace the rough ratios of the first Basel accord were found to add, rather than to reduce, the procyclicality of bank behavior. This is because they were themselves based on market valuations of assets and encouraged banks to expand their balance sheets in good times and to shrink them in bad times.

In the US, risk-weighted capital ratios were supplemented with a cruder, non-risk-weighted capital-to-asset ratio called a *leverage ratio**. Major bank holding companies (not investment banks) were required to hold Tier-1 capital (see box 8.3) of at least 4% of their total assets.[13] The US leverage ratio did not prevent the crisis but it may be the case, as argued by the Swiss vice-governor Philipp Hildebrand (2008), that "it ensures a minimal buffer to absorb the negative consequences of imprudent behavior." World leaders decided at the G20 Pittsburgh summit to incorporate it into the Basel II framework as a supplement to the capital-adequacy ratio, and to make it compulsory when valuation standards have converged sufficiently so that the denominator of the ratio is measured consistently across countries.

Box 8.3 Why Are Capital-Adequacy Ratios Procyclical?

The setting of minimal capital requirements is intended to provide a buffer so that banks remain solvent across a wide range of shocks. As such, it is an application of the Value-at-Risk approach examined in chapter 2.

13. The link between the capital-to-asset and debt-to-asset ratios can be understood using the above notation: $K/A = 1/(1 + l)$.

*Capital-adequacy ratios** for internationally active banks were first introduced in 1988 by the Basel I accord, which imposed on banks a minimum capital of 8% of risk-weighted assets. Risk was supposed to depend on the asset class, so for example governments were deemed safe and corporate bonds risky. In the 1990s this crude approach was subject to criticism, which resulted in the introduction in the mid-2000s of the new Basel II ratios.

Basel II introduced two main innovations. First, two categories of capital were distinguished: *Tier 1 capital**, which broadly corresponds to shareholder equity, and Tier 2 capital, consisting of reserves, provisions and subordinated debt. The 8% ratio has as numerator total Tier 1 + Tier 2 capital, with the proviso that Tier 2 capital must be inferior or equal to Tier 1 capital. Second, Basel II differed from Basel I in its approach to risk, which is not given for broad asset classes anymore but is asset-dependent and time-varying. It can be calculated according to one of two methods, the *standardized approach* and the *internal ratings-based approach.*

- The standardized approach uses ratings published by the credit rating agencies to measure risk. Both loans to governments and loans to corporations are therefore deemed risky and enter into the calculation of total credit risk, with weights dependent on their ratings (for example, in the original Basel II framework claims on governments rated AAA did not enter into the calculation of total risk, while claims on governments rated BBB were taken into account for 50% of their value; for claims on corporations, the corresponding weights were 20% for AAA borrowers and 100% for BBB borrowers). Once the total risk has been calculated, the minimum capital adequacy ratio (8% in the original Basel II framework) is applied to determine the bank's minimum capital.
- Alternatively, banks can be authorized by their supervisor to use an internal ratings-based approach, whereby weights are determined by the bank's own assessment of the riskiness of its claims on the basis of methodologies and parameters determined by the regulator. For example, the original Basel II framework required banks to compute the maximum losses that they could suffer at a 99.9% confidence interval. The bank would be required to hold at least enough capital to absorb this "maximum probable" loss. However, the evaluation of a borrower's probability of default was left to the bank itself. The intention was to make better use of a bank's internal information on the riskiness of its clients and to better take into account the correlation of risks across assets within the bank's portfolio.

The standardized approach is subject to procyclicality to the extent that credit ratings are themselves procyclical, which tends to be the case although rating agencies claim to smooth risk assessment over the cycle.

Simple empirical evidence indicates that average ratings decline in a downturn, leading to an increase in capital requirements. Similarly they improve in boom times, relaxing capital requirements (Panetta and Angelini, 2009). Instead of dampening the procyclical effects of leverage, regulation therefore tends to increase them.

There is no direct empirical evidence yet on the internal ratings-based approach since it was introduced in 2008 only (and only in Europe) but simulations have shown that it is likely to be open to the same criticism as the standard approach. For example Repullo and Suarez (2008) find that the Basel II framework provides better protection against bankruptcy than Basel I but that, since banks are unlikely to hold sufficient buffers above the minimum requirements, the increased risk of borrower default during a recession should imply credit contraction in downturns.

d) Market valuation

It has been noted above that leverage leads banks to respond procyclically to fluctuations in the value of their assets. This raises the question of how bank assets should be accounted for, which is a complex and as-yet unresolved question.

In the years before the crisis, the financial reporting of banks had been increasingly based on so-called *fair-value** accounting:[14] Assets and liabilities were reported at market value, with capital gains and losses being registered in the profit-and-loss account. When market prices were not available, fair value was constructed by discounting expected future cash flows, based on some forecasting model. This principle was enshrined in the *International Financial Reporting Standards* (*IFRS*)*, adopted by more than 100 countries including EU member states, and in the *US Generally Accepted Accounting Principles* (*GAAP*)*.

There are several issues with fair-value accounting:

- *Consistency between standards*: A given asset may be valued differently by the bank's supervisor and by its auditors, and both standards may vary across countries. Deutsche Bank, a German bank with significant US activities, reported total assets worth 2202 billion dollars under IFRS and 1030 billion dollars under US GAAP as at 31 December 2008. This is because financial derivatives are registered at gross value under IFRS and at net value under US GAAP.
- *Availability of market prices*: The crisis has led accounting standard-setters to acknowledge (somewhat reluctantly, at least initially)

14. Fair value accounting is a broader concept than *mark-to-market** accounting. It allows reliance on other methods, such as the use of models-based valuation when there is no market price to base the valuation on.

that market valuation is not possible when markets do not function. It may remain possible to discount expected future cash flows but investors are suspicious of biases in prices produced by fragile and potentially self-serving internal models.

- *Counter-intuitive outcomes*: For example, when the creditworthiness of a bank deteriorates, the market value of its liabilities goes down and it can therefore register a profit in its profit-and-loss account.
- *Procyclicality:* A fall in asset prices induces banks to sell assets and contract credit in order to comply with capital requirements. Box 8.4 provides a telling illustration in the case of pension funds: The combination of a strict pension funding rule and mark-to-market accounting produces an upward-sloping demand curve on asset markets: When the price of bonds goes up, pension funds have to buy more of them. Such behaviors exacerbate disruptive market dynamics.

There is a minority view that market valuation should be abandoned altogether in favor of historical cost valuation, or strictly limited to trading activities.[15] Based on the experience of past financial crises (particularly the Japanese one), the economic profession generally considers that this would obscure the perception of banks' soundness, delay the necessary disposal of non-performing assets, and eventually aggravate the cost of crises. An alternative is to supplement mark-to-market accounting with appropriate clauses so as to mitigate its procyclicality, such as buffers to weather sudden drops in market prices, and temporary waivers in case of a crisis.[16]

Box 8.4 Procyclical Mark-to-Market Accounting: The Case of Pension Funds

We illustrate here how mark-to-market accounting may force financial institutions to act in a procyclical way on financial markets. The example is adapted from Boeri et al. (2006).

Consider a pension fund with pension disbursements l_t at all future dates $t \geq 0$. For analytical convenience, we suppose the fund is entirely invested in perpetual bonds with a unitary face value, yielding a constant interest rate r. The market value of the bond portfolio is $A = pN$, where N is the number of bonds and $p = 1/r$ is their unit price. The model is written in continuous time.

Looking forward, pension liabilities increase at a constant rate λ (say, because pensions are indexed on wage growth): $l_t = l_0 e^{\lambda t}$ with $0 < \lambda < r$.

15. In 2003, French President Jacques Chirac wrote to European Commission President Romano Prodi that the adoption of fair-value accounting would "lead to company management methods that will place excessive bias on the short term."

16. In response to the crisis, the US Financial Accounting Standards Board (FASB) and the International Accounting Standards Board (IASB) authorized the temporary valuation of some assets at historical value rather than market value in October 2008.

Since there is no active market for pension portfolios, their fair value L is computed using discounted expected cash flows:

$$L = \int_{t=0}^{\infty} (1+r)^{-t} l_t = \frac{1}{r-\lambda} l_0 \qquad \text{(B8.4.1)}$$

Let $\varphi = A/L$ be the *funding gap** of the pension fund, i.e., the discrepancy between its market-valued assets and liabilities. When interest rates go down (or, equivalently, when the price of bonds goes up), the value of liabilities increases more than the value of assets and the funding gap widens:

$$\frac{1}{\varphi}\frac{\partial \varphi}{\partial r} = -\left(\frac{1}{r} - \frac{1}{r-\lambda}\right) > 0 \qquad \text{(B8.4.2)}$$

Suppose now that the price of bonds p fluctuates and the fund manager, facing a given liability portfolio, adjusts in real time the size N of the asset portfolio to match a given funding gap φ (say, as imposed by pension fund regulation). The manager's rule is:

$$N(p) = \frac{A(p)}{p} = \varphi \frac{L(p)}{p} = \varphi \frac{1}{1-\lambda p} l_0 \qquad \text{(B8.4.3)}$$

We get:

$$\frac{1}{N}\frac{\partial N}{\partial p} = \frac{\lambda}{1-\lambda p} > 0 \qquad \text{(B8.4.4)}$$

Under the combination of a regulatory funding rule and mark-to-market accounting, the fund has to buy *more* bonds when their price goes up. When applied to the whole industry, such rules may exert a destabilizing, procyclical impact on bond markets. This impact was documented on the euro and sterling bond markets when pension-fund regulation was tightened and moved to mark-to-market valuation in Scandinavia, then in the UK in the early 2000s (Boeri et al., 2006).

e) Why did the subprime crisis trigger a generalized panic?

It is now time to answer our second question, i.e., why a crisis in a limited segment of financial markets, namely the subprime market, contaminated the entire financial system. According to the IMF (2010a), writedowns on mortgage-based securities incurred by US banks in the 2007–09 crisis amounted to some $200bn or less than 1.5% of GDP, a not-insignificant amount but a much lower one than the losses recorded by the savings and loans institutions in the early 1990s. Adding losses on nonsecuritized loans or losses incurred by non-US banks increases the absolute amounts involved but does not change the conclusion: The financial meltdown cannot primarily be ascribed to the weight of subprime securities in investors' portfolios.

Furthermore, as developed by Gorton (2008, 2009b), special investment vehicles had broadly diversified portfolios and were not significantly exposed to subprime loans.[17]

A key element to understanding how a localized crisis became a global one was the contamination through assets held as collateral in the market for *repurchase agreements* or *repos* (see box 8.5). In this market, lending firms deposited cash, borrowing firms posted securities as collateral, and this collateral could in turn be "rehypothecated" in exchange for cash with a third party. Gorton (2009b) considers that this market and the "shadow banking system" that underpinned it fulfilled a role analogous to that of a banking system for financial institutions, because it allowed these institutions to deposit cash and borrow without being exposed to counterparty risk.

Gorton points out that the essence of banking is the provision of liquidity through producing what he calls "informationally-insensitive debt": Thanks to deposit insurance, which prevents bank runs (see chapter 4), demand deposits are regarded as as good as central bank money and no one can derive any profit from the production of private information about them. Securitization was a way to create "relatively informationally-insensitive debt" without deposit insurance. The posting of securitized assets as collateral provided the means to meet the borrowing needs of some firms and the demand for liquid, informationally insensitive deposits of some other firms. However, it also increased the complexity so that information on the distribution of risks was scarcely available and increasingly costly to assemble. It also resulted in an exponentially increasing demand for safe assets to be used as collateral—we will come back to this point later.

Box 8.5 Repo Transactions

Repurchase agreements, or *repos**, are short-term loans backed by an exchange of collateral (see also the description of central bank repo transactions in chapter 4).

In this market, *counterparty risk** (i.e., the risk that the credit extended is never recovered) is only residual provided that the amount of collateral is revised frequently enough to offset the change in value of the asset deposited as collateral. This is usually done through cash deposits called *margin calls**. Typically, Bank A borrows X million dollars from Bank B for a given, short period of time and transfers to Bank B for the life of the loan a pool of assets worth the same amount. Bank B then regularly checks the market value of these assets. If they have depreciated by x%, Bank A

17. Gorton (2009) also questions the relevance of an explanation based on the "originate and distribute" view, according to which risks were passed along to investors, thus lessening incentives to care about risk. He argues that risks remained all along the chain from originators to underwriters, and that the interests of the various parties were aligned in securitization.

transfers to Bank B an additional xX million dollars in cash as a margin call. This ensures that bank B's loss will be limited if Bank A defaults. An alternative, which can be combined with margin calls, is to impose an arbitrary rebate on the value of collateral (called a *haircut**), depending on its creditworthiness. This increases the quantity of collateral required in exchange of a given monetary amount.

Repos involve less counterparty risk than uncollateralized bank loans and have therefore developed very rapidly. While there are few statistics about this market, it was believed to exceed 10 trillion dollars in 2008, having grown by around 10% a year (Gorton, 2009a). This amount was roughly equivalent to the total assets of the banking sector.

As long as the system expanded steadily, no question needed to be raised about the quality of collateral. However, the leveraging and tranching mechanism implied that the price of a subprime asset-based security used as collateral was a highly nonlinear function of house prices. In spite of the moderate share of subprime bonds in the pool of asset-based securities, the bursting of the real-estate bubble affected the valuation of collateral and thus transformed what was perceived as "informationally insensitive" debt into "informationally sensitive" debt.

The complexity of the whole chain of structured financial products meant that the information necessary to properly value claims was not accessible. No one could accept structured products as collateral any longer. The subprime crisis thus translated into a collateral crisis and a dash for cash. Depositors were not able to assess counterparty risk. Average repo haircuts exploded, from zero in the first half of 2007 to 25% by mid 2008, and more than 45% by the end of 2008. The repo market dried up. The demand for cash could only be met by selling assets at much reduced prices, so that the price of nonsubprime related assets also fell substantially. The mark-to-market value of all assets collapsed, feeding back into a further drying up of the repo market and solvency problems for financial intermediaries. The failure of Lehman further compounded both the signal, the dash for cash, and the panic. There was, in Gorton's words, a "run on the shadow banking system."

As suggested by Holmström (2008), this information-asymmetry problem was not primarily one of transparency, but rather one of complexity. The whole system thrived on non-transparent information, and it is when price information became more collective and transparent that the panic unfolded.

8.1.5 Macro roots

"At the core of the crisis lay an interplay between macro-imbalances which had grown rapidly in the last ten years, and financial market developments

and innovations." The gist of this sentence, from the Turner Review (Financial Services Authority, 2009, p. 11) commissioned by the UK government, can be found in many other assessments by experts and, interestingly, regulators.[18] Beyond the microeconomic roots and the failures of regulation, broader permissive factors were conducive to financial imprudence.

In fact, if interest rates had been higher, housing booms, stock market valuations, and the rise in private debt would certainly not have reached the same levels. Cheap credit facilitated debt-financed investment in real estate and financial assets, and contributed to excessive risk-taking. From a macroeconomic standpoint also, this crisis has been a crisis of leverage (Figure 8.2).

Almost by definition, macroeconomic factors therefore played a role in the boom–bust cycle, because interest rates affect the demand for credit: There is necessarily an interest-rate level that would have prevented the boom. But the interesting question is: What created this macroeconomic environment? Was it a failure of monetary policy? Was the broader saving–investment balance at global level the root cause of the low interest rates it produced? Although mutually compatible, these two explanations have quite different policy implications.

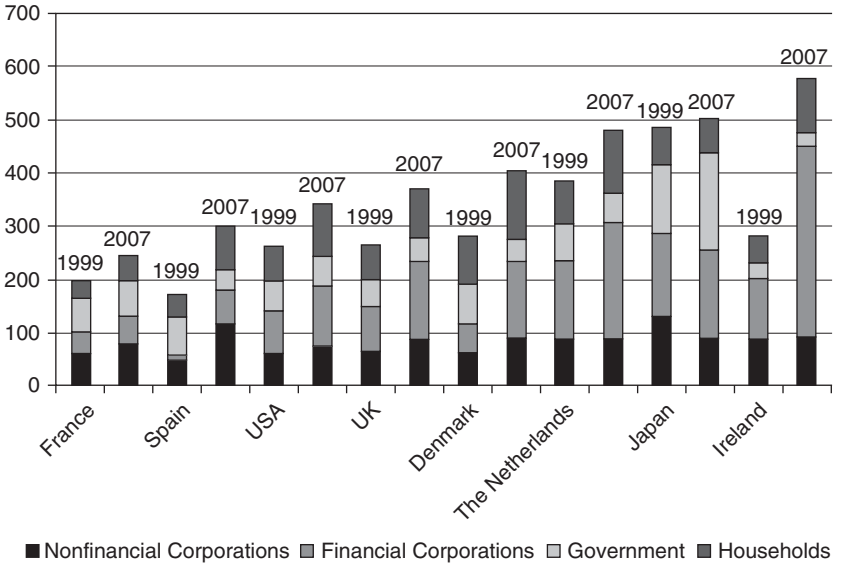


Figure 8.2 The rise of private debt between 1999 and 2007. Interest-rate liabilities (loans and nonequity securities) as % of GDP.
Sources: Eurostat, ECB, Federal Reserve and Barclays Capital.

18. See for example the De Larosière (2009) report prepared at the request of the European Commission.

a) A failure of monetary policy?

A first explanation blames an exceedingly lax monetary policy, either in the US (Taylor, 2008) or globally (Bank for International Settlements, 2008). According to this view, monetary policy in the aftermath of the 2001 recession remained too lax for too long and this triggered both asset-price inflation, primarily but not exclusively on the US housing market, and a generalized leverage boom.

Figure 8.3 depicts the evolution of policy interest rates (the Fed Fund rates) and of 10-year Treasury interest rates from the late 1990s to the late 2000s. The dashed line, taken from Taylor (2009), represents the counterfactual Fed Funds evolution that would have been observed had the central bank followed a Taylor rule.[19] The Fed would have tightened rates faster after the 2001 recession, instead of lowering interest rates further to counter perceived deflation risks. Accordingly, short-term rates would have been higher between 2001 and 2005, denting the housing price boom and making the subsequent bust less pronounced.

In retrospect, the Fed should have worried less about the deflation risk in 2003, when then-board member Ben Bernanke famously outlined a contingency plan to avoid the repetition of the Japanese experience (Bernanke, 2002), and it should have worried more about the risks of a housing bubble,

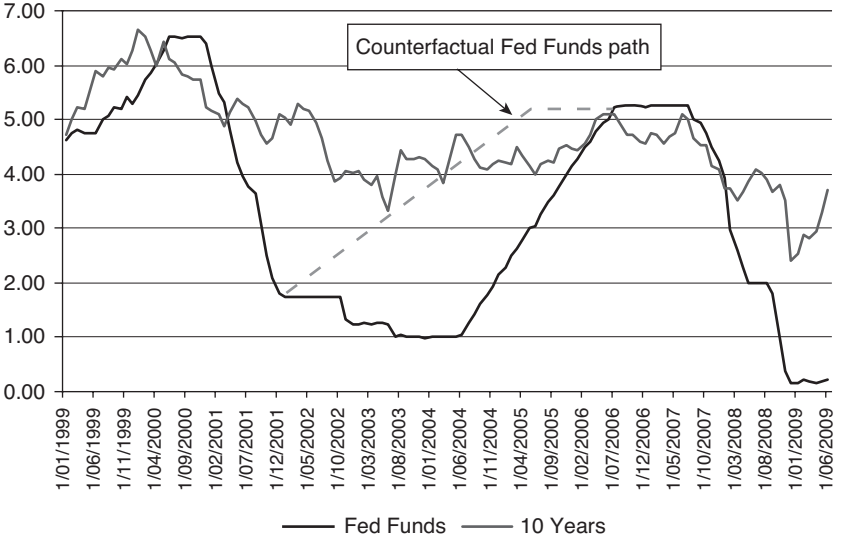


Figure 8.3 US interest rates, 1999–2009.
Source: Federal Reserve Bank of St Louis.

19. The Taylor rule is presented in chapter 4.

instead of claiming, as then-chairman Alan Greenspan did, that "while local economies may experience significant speculative price imbalances, a national severe [housing] price distortion seems most unlikely in the United States." (Greenspan, 2004). This illustrates the difficult art of risk-management in economic policymaking: Faced with these two equally improbable outcomes, the Fed may have overstated the former and disregarded the latter.[20]

The question, however, is whether this explanation is *sufficient*:

- To start with, the Taylor rule can only give rough indications and cannot be taken as an undisputable benchmark. Consumer-price inflation remained rather subdued throughout the 2000–06 period and accelerated only with the world commodity-price boom of 2007–08. A reason for continued price stability was the flattening of the Phillips curve discussed in chapter 4. Central-bank credibility, structural changes in the US labor market, and the increase in the global labor force resulting from China's and India's increased participation in globalization all resulted in a containment of wage and price increases. A central bank dedicated to price stability (rather than to a combination of inflation and the output gap, as in the Taylor rule) had therefore little reason to raise interest rates aggressively enough to prick the real-estate bubble;
- The question is rather, therefore, whether the Fed should have raised interest rates *in the name of financial stability*. John Taylor implicitly assumes that by following a Taylor rule it would have killed two birds with one stone—achieving both macroeconomic and financial stability. However, there is no theoretical or empirical motive to believe that the two objectives are coincident. As discussed in chapter 4, whether central banks should explicitly target asset prices when setting interest rates has been a matter for debate. Put simply, it implies that central banks stand ready to depart from their macroeconomic stability goal in the name of financial stability—not something they can consider lightly. On a more practical ground, whether the Fed could have steered interest rates delicately enough to engineer a soft landing of housing prices is dubious;
- Furthermore, from 2001 on, long-term interest rates remained remarkably stable at a low level (figure 8.3) consistent with stable inflationary expectations. This stability, famously dubbed a "conundrum" by Alan Greenspan (2005), contrasted with previous episodes when bond rates responded to movement in policy rates, and it suggests that, if the Fed had followed John Taylor's *ex post* prescription, bond rates could have remained at a low level. This leads us to consider the structural reasons for the persistence of low long-term interest rates throughout the early 2000s.

20. Bernanke (2010) disputes John Taylor's critique of the Fed behavior in the 2000s on the grounds that even assuming it should have followed a Taylor rule, real-time statistical information and forecast did not warrant the policy reaction depicted in figure 8.3.

b) A consequence of global imbalances?

An alternative macroeconomic explanation for the crisis focuses on global current-account imbalances rather than on purely domestic developments. According to this view, the increased demand for safe assets associated with capital flows into the US favored leverage and even provided incentives to manufacture purportedly AAA assets of actually dubious quality.

The starting point for this analysis is the observation of a massive inflow of foreign savings into the US. As the US came out of the 2001 recession, a new global saving–investment pattern emerged that characterized the 2002–07 period.[21] What became known as "global imbalances" was the combination of an historically high, and growing US current-account deficit of the order of 1.5% of world GDP (average over the 2002–07 period), and corresponding surpluses in East Asia and later in the oil-producing countries. During this period foreign net purchases of US Treasury securities always represented more than 60% of net issues and for the entire period they amounted to 81% of total net issues.[22]

In the early 2000s, the traditional view of global imbalances—a view generally endorsed by Europeans—was that they were primarily driven by the US saving behavior—in other words that they resulted from a domestically rooted drop in the US saving rate. In the mid-2000s, however, Ben Bernanke put forward an alternative explanation, deemed the "global savings glut" hypothesis (Bernanke, 2005). According to this view, the global savings–investment pattern originated in an increase in the rest of the world's net saving, rather than resulting primarily from US behavior.

However self-serving for the US, Bernanke's provocative thesis rightly pointed out that financial globalization and the appetite of emerging countries for US Treasury bonds had to feature in the analysis of global imbalances.[23] The question became: Why was the rest of the world so keen on investing in US assets? Three main rationalizations have been offered for such behavior: The asset shortage, self-insurance, and Bretton Woods 2 hypotheses.

The *asset-shortage hypothesis** presented in Caballero et al. (2008) and Mendoza et al. (2009) posits that financial underdevelopment in emerging countries led domestic agents to export their savings and invest them in (US) assets of higher safety and quality. This simultaneously resulted in a US financial account surplus (and a corresponding current account deficit) and in a lowering of long-term interest rates, as foreign savings increased the demand for financial assets. This intellectually attractive explanation has, however, not been tested extensively. Furthermore, the asset-shortage hypothesis does not

21. As documented elsewhere in this book, the US current-account deficit goes back in fact to the very early 1980s. It had been reduced by the turn of the 1990s before deepening sharply in the second half of the decade.

22. Data here are taken respectively from the IMF's World Economic Outlook and the US Flow of Funds statistics.

23. See also chapter 5.

explain why the emerging countries' investment in the US overwhelmingly came from central banks: According to the IMF's COFER data (which are far from exhaustive as many central banks do not disclose the allocation of their reserves), the developing and emerging countries' dollar reserves rose from US$258bn in 1999 Q1 (in the immediate aftermath of the Asian crisis), to US$2254bn in 2008 Q2 (before the turmoil that followed the Lehman collapse), an almost 10-fold increase in less than 10 years.

One rationale for such accumulation was to avoid a repetition of the 1997 balance-of-payment Asian crisis and subsequent dependence on IMF financings, perceived as costly and humiliating: Instead, international reserves were used as *self-insurance** against future crises. This rationalization, however, is not entirely satisfactory either, especially concerning China: Self-insurance may explain a one-off increase in foreign-exchange reserves but the continuous accumulation of low-yielding reserves involves a significant opportunity cost that is hard to justify from a social planner's point of view (Rodrik, 2006).

Another rationale, especially in China, was the export-oriented growth strategy that implied keeping the currency undervalued through repeated interventions on the foreign-exchange market. Dooley et al. (2003) spoke of a "*Bretton Woods 2 regime*" (see chapter 5) to describe the resulting web of explicit or implicit exchange-rate arrangements between the dollar and the developing and emerging countries' currencies. The bulk of corresponding central-bank reserves were held in US Treasury bonds because they were the most liquid (and supposedly the safest) securities in the world and because the currencies were *de facto* pegged to the dollar.

Whether or not global imbalances were sustainable has been a matter for fierce debate within the economic profession. For some scholars (e.g., Engel and Rogers, 2006) the US current-account deficit was the perfectly natural result of intertemporal optimization by US consumers, while for others (Obstfeld and Rogoff, 2005; Blanchard et al., 2005), it was unsustainable in the long run. The latter, however, generally expected a precipitous decline in the US-dollar exchange rate, possibly accompanied by a sell-off of US government bonds, not a *domestic* financial crisis.

After the event, the crisis revealed an unanticipated link between the foreign search for safe assets and US domestic risk-taking. Intuitively, the low level of long-term rates resulting from capital inflows led investors from the US and other industrialized countries to diversify away from "plain-vanilla" US Treasury securities and look for higher-yield paper, thereby encouraging the manufacturing of securities that were granted AAA status by rating agencies but which offered a higher return than Treasury bonds. CDOs, or at least the degree of success of CDOs, were the product of this link. The US was playing its traditional role as the "world venture capitalist,"[24] borrowing from risk-adverse Asian investors and investing into risky assets.

24. This expression is borrowed from Pierre-Olivier Gourinchas and Hélène Rey (2007).

However, these were no longer productive investments but toxic leveraged products. Caballero and Krishnamurthy (2009) provide a simple model of such a link between global imbalances and US financial fragility and show how foreign demand for safe US assets could contribute to the rise of leverage and the fall in risk premiums (box 8.6).[25]

The link between global imbalances, low long-term interest rates, leverage, and the supply of seemingly safe financial products has, however, not been documented empirically. Blanchard and Milesi-Ferretti (2009) dispute it, at least implicitly. Instead of putting the emphasis on net savings flows (as the global-imbalances approach does), they prefer instead to emphasize the role of *gross* cross-border holdings of financial assets in the transmission of the crisis from the US to Europe. Both explanations, however, are compatible. Linkages between global imbalances, low long-term interest rates, leverage, and the development of new financial products have not yet been assessed systematically. Warnock and Warnock (2009) explore the impact of foreign official capital inflows on US long-term interest rates and find that they may have depressed them by close to 100 basis points in 2005, which is not a trivial effect. In a broader perspective, Obstfeld and Rogoff (2009) discuss the impact of low interest rates on financial innovation and claim that global imbalances and the crisis had common causes.

Box 8.6 Global Imbalances and US Financial Fragility: A Simple Model

The model, adapted from Caballero and Krishnamurthy (2009), has three agents: Domestic financial firms, domestic investors, and foreign investors.

Domestic financial firms generate a cash flow X_t per unit of time that comes from their portfolios of loans, e.g., mortgages. Let V_t be the present value of these future cash flows. The financial firms are leveraged and issue debt to the amount of B_t. The debt is deemed safe and pays the risk-free interest rate $\boldsymbol{r}$. The equity value of the financial firms is therefore:

$$W_t = V_t - B_t \tag{B8.6.1}$$

Domestic investors hold financial firms' equity and their wealth is therefore W_t. They consume a fixed fraction ρW_t of their wealth per unit of time, in conformity with a behavior optimally derived from log preferences.

Foreign investors are more risk-adverse and hold only debt B_t (think of foreign central bank holdings). They invest a flow X_t^* and repatriate a fraction $\rho^* B_t$ of their wealth per unit of time.

25. See Caballero (2009) and Brender and Pisani (2009) for developments along these lines.

These are crude assumptions intended to capture the behavior of US and foreign emerging countries' investors in the 2000s. It would not change the results to assume that the two categories of investors hold both equity and debt as long as they have a different preference for the two categories of assets.

The goods market equilibrium is written as:

$$\rho W_t = X_t + X_t^* - \rho^* B_t \tag{B8.6.2}$$

That is, domestic consumption equals the debt stream from financial firms plus net capital inflows. This equation can be solved for the equity value of domestic financial firms:

$$V_t = \frac{X_t + X_t^*}{\rho} + \left(1 - \frac{\rho^*}{\rho}\right) B_t \tag{B8.6.3}$$

The first term on the right-hand side indicates that foreign demand for *riskless* assets increases the equity value of financial firms, i.e., of domestic *risky* assets (and therefore the wealth of domestic residents W). This is because leverage brought about by the foreign demand for safe assets increases the value of equity. The second term indicates that the increase is stronger if foreign asset-holders have a lower propensity to consume (repatriate) their wealth than domestic asset-holders.

In the same way it can be shown that if capital inflows are stable, then the foreign demand for safe assets lowers the risk premium on domestic risky assets.

8.1.6 The "Black Swan" syndrome

Complex systems are prone to accidents and the more integrated they are, the more catastrophic the accident can be. Financial markets are specialized in dealing with risk but are not prepared to face extreme events. When such events materialize, the whole system may collapse. "Complexity got the better of us," wrote Goldman Sachs CEO Lloyd Blankfein in February 2009, adding that we should resist a response, however, that is solely designed around protecting us from the 100-year storm because "taking risk completely out of the system will be at the cost of economic growth" (Blankfein, 2009).

Very few observers, if any, go so far as saying that the crisis was purely a "Black Swan," i.e., a large-impact, low-probability event against which any protection would have been exceedingly costly.[26] But many give it a certain weight and use it to caution against the temptation to overprotect. It is also a challenging intellectual hypothesis that deserves to be explored.

26. The black-swan metaphor is attributed to Nassim Nicolas Taleb (2007) and has its root in the observation by Karl Popper, the twentieth century philosopher, that seeing no black swans was not a proof that black swans did not exist.

As already observed, the subprime crisis was in itself a relatively minor event. According to the International Monetary Fund (2008), the losses on US nonprime mortgage loans that set in motion the dramatic chain of crisis events stood in October 2008 at some 100 billion dollars. This corresponded to just 0.7% of US GDP and 0.2% of world GDP, a small amount in comparison to eventual, global losses. Similar comparisons could be made with the emerging markets crises of the 1990s. Yet the consequences of the previous episodes remained contained.

We have explained what role the use of securitized assets as collateral has played in the transmission of the shock. However, the issue runs deeper. Andrew Haldane, the Bank of England's director for financial stability, has drawn interesting comparisons with collapse phenomena affecting other complex, network-based systems such as electricity grids and complex ecosystems, for example rainforests or fish stocks. Such systems exhibit strong nonlinearities in response to shocks and, according to Haldane, they are at the same time both robust and fragile. Their complexity and connectivity makes them resilient to a wide range of shocks because "the system acts as a mutual insurance device with disturbances dispersed and dissipated. Connectivity engenders robustness. Risk-sharing—diversification—prevails. But beyond a certain range, the system can flip the wrong side of the knife-edge. Interconnections serve as shock-amplifiers, not dampeners, as losses cascade. The system acts not as a mutual insurance device but as a mutual incendiary device" (Haldane, 2009).

There is strong evidence that the very strategies that were intended to limit risk—especially securitization and insurance through derivative products—dramatically increased the complexity of the financial system and at the same time reduced its diversity, because all firms were following similar strategies and were making themselves vulnerable to the same events. Such lack of diversity can explain why a relatively small shock became so greatly amplified through the financial system. If this interpretation is correct, the black swan may show up again in the future: Instead of being an unpredictable, once-in-a-century event, big crises are an endogenous property of a robust-yet-fragile system in the same way that collapses are an endogenous property of the robust-yet-fragile integrated electricity grids. If this is the case, responses should focus not on checking whether each and every part of the system is in good shape but on improving the stability of the whole. This may imply *stress-testing** the financial system, i.e., assessing the impact on banks' balance sheets of various scenarios involving the propagation of shocks across the financial system, and protecting vital elements of the financial system from the contagion of its riskier segments—as was done after the Great Depression with the introduction of the Glass–Steagall Act of 1932 that separated investment banking from commercial banking—or giving to a specific institution the mandate to oversee global financial stability, over and above the mission industry regulators are entrusted with. We return to these issues in section 8.3.

8.1.7 Lessons

It would be pointless to try to determine which of the three approaches to the crisis reviewed above is the most relevant, or even to establish a hierarchy between these different sets of potential causes. The reason is that they touch upon different policy domains—financial regulation, monetary policy, international coordination—and are mutually reinforcing. For instance, excess leverage due to insufficient regulation was encouraged by low interest rates. Why were interest rates so low? The Fed's monetary policy provides an immediate answer. However, without the international appetite for US Treasuries, the US dollar would have been weaker, triggering import-price inflation and forcing the Fed to increase interest rates. And at longer time horizons, interest rates are determined by international capital markets rather than by local monetary policy. More directly, the global demand for US dollars spurred the production of dollar-denominated assets.

More generally, through its basic ingredients, this crisis resembles previous crises experienced throughout history: Asset-price bubbles financed through leveraging, followed by a market scramble. Why, then, was the crisis not anticipated? The reason is twofold.

First, as argued above, the roots of the crisis are to be found in different spheres. Robert Shiller of Yale warned against the risks of a US housing-price collapse. The IMF repeatedly pointed out the burgeoning US current-account deficit, and Maurice Obstfeld of Berkeley, Kenneth Rogoff of Harvard, and Nouriel Roubini of New York University, among others, anticipated a dollar crisis. Michel Aglietta of Paris-Nanterre and Claudio Borio of the BIS warned policymakers against systemic risk developing in the banking sector.[27] However, few economists were able to embrace all dimensions of the crisis, from accounting and banking standards to global current-account imbalances, from the intricacy of ABS markets to off-balance banking conduits.

Second, after an unprecedented period of expansion, and a succession of eventually benign financial crises, the crisis found policymakers and their advisors sleeping at the wheel. The crisis of the junk-bond market in 1989, the demise of LTCM in 1998, the bursting of the dot-com bubble in 2001 were all significant events in the financial sphere, but none of them resulted in a world recession. This created confidence in the robustness of the system and a sense of complacency, which was proved wrong by the 2007–08 crisis. More globally, crisis prevention faces the well-known hurdles of collective action: The change of behavior that is necessary to heed the various signals that are always available not only requires individual wisdom but makes sense only if it is implemented collectively.

27. See Obstfeld and Rogoff, 2009, for an account of the various stages of the pre-crisis discussion.

8.2 Extraordinary Times

In this book, we have presented the evolution of policy thinking and policymaking in the post-war period. As developed in chapter 1, a clear pattern emerging from this evolution has been a guarded approach to government intervention. In the mid 2000s, virtually any minister, central banker, or regulator in the world contemplating policy action started off by asking himself or herself whether public intervention was necessary and whether it would risk doing more harm than good. Even those who (like the authors of this book) did not share a belief in the self-regulating character of markets, acknowledged that government failures were probably as pervasive as market failures (some of the reasons are discussed in chapter 2) and that before embarking on public intervention a thorough examination of the pros and cons was needed.

Another, related, pattern of policymaking has been the increased emphasis on predictability. As developed in chapters 2 to 5, in accordance with the rational-expectation paradigm, economic policy came to be seen in late twentieth century as a repeated game against intelligent players. The consequence was to lay stress on the clarity of objectives and the growing importance of policy rules—even when rules were intended to serve as benchmarks rather than strict guidelines. Examples of such rules were the budgetary rules introduced in chapter 3 and the monetary rules introduced in chapter 4, but the same approach was also extended to other areas. This pattern was especially apparent in Europe where policy by rules was enshrined in the EU treaty.

The lessons of the crisis for economic policymaking in normal times will be discussed in section 8.3. Crisis management, however, calls for a different kind of policy behavior. In the same way wartime governance departs from peacetime governance, it involves actions that break with the traditional boundaries between private and public domains and disregard rules-based guideposts. Instead of predictability it requires speed of action, flexibility, and innovation. It thus brings policymaking onto entirely new territory where the usual compass is of little use beyond drawing attention to the inevitable day of reckoning when the full costs of heterodoxy will need to be dealt with. This section is about this new territory.

8.2.1 Economic policy without the usual compass

In August 2007 central banks embarked on providing wholesale liquidity to financial institutions—not knowing, at the time, how far the journey would take them. In October 2008, governments came to the rescue of ailing banks in order to avoid further bankruptcies and to revive the credit channel. Simultaneously, central banks lowered interest rates aggressively, soon reaching the zero bound, while fiscal policy turned expansionary.

a) The rescue of ailing banks

From the 1980s until the mid 2000s, privatization had been a policy mantra in both developed and developing countries. Empirical research had supported the proposition that privately owned firms are more efficient and more profitable than otherwise-comparable state-owned firms (Megginson and Netter, 2001). Unless there was a clearly stated general interest argument, public ownership of commercial banks or nonfinancial companies was regarded as evidence of a lack of clear policy objectives, and was even considered as a handicap, as it confronted policymakers with a conflict of interest between their role as shareholders and their role as regulators. Either banks benefited from privileged access to government support, which raised competition concerns, or they had to behave like private banks, which deprived public ownership of any purpose. In most countries consequences were drawn: The public banking sector was limited to general-interest institutions such as development banks, and when it survived its privileges were eventually sacrificed on the altar of competition, such as the state guarantee of the borrowing of the German *Landesbanken**.[28]

In 2008, however, governments in most countries had hurriedly to reverse this stance and found themselves doing the opposite of what they had claimed was their doctrine. Capital injections into banks amounted in most cases to several percentage points of GDP (up to 6.5% in the UK and roughly as much in smaller European countries like The Netherlands, Belgium, and Ireland), either through outright participation and control of the bank, or by subscribing to preferred shares (see box 8.7) to avoid taking control.

When a large bank is unable to roll over its debt in spite of short-term liquidity provision by the central bank and faces a threat of failure, the government can either let the bank fail—and possibly face the systemic consequences; or it can save the bank and in the process bail out its depositors and lenders, thereby creating moral hazard. Bank failures are not exceptional events. In 2008, 26 US deposit banks were allowed to fail, a number still small in historical terms, but in 2009 as many as 140 banks filed for bankruptcy (figure 8.4).[29] However, most failed banks were small enough not to trigger a domino effect in the banking system.

Lehman Brothers' failure could have been a salutary reminder to all holders of bank shares of the risk associated with the high returns on their holdings, thereby helping to keep moral hazard in check. In fact, it turned to disaster

28. The regional banks (*Landesbanken*) were forced by the European Commission to abandon their state guarantees because these represented a distortion of competition. This termination resulted in a borrowing spree before the expiration of the guarantee in 2005, and its proceeds gave rise to hazardous investment in high-yielding assets such as US mortgage-based securities.

29. More than 500 deposit banks failed in 1989, and up to 4000 in 1934. See Gorton (2009b).

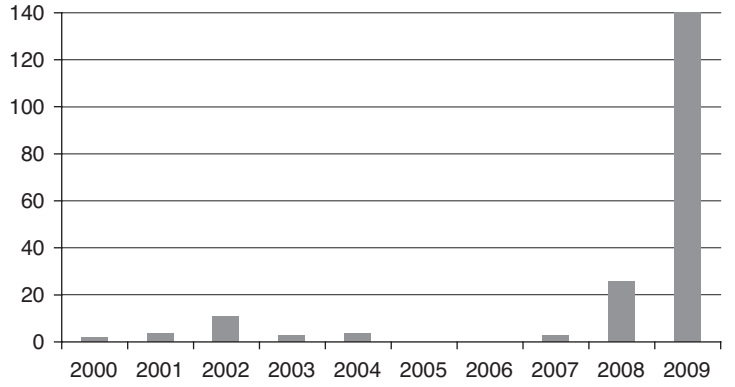


Figure 8.4 Bank failures in the US, 2000–2009.
Source: FDIC.

due to the size and interconnectedness of Lehman Brothers, and because it triggered a massive loss of confidence. Only three weeks after Lehman's failure, on 10 October 2008, G7 finance ministers announced an unequivocal change of course, saying that they would "use all available tools to support systemically important financial institutions and prevent their failure."

The question had moved from *whether* to intervene to *how* to intervene. In this respect, past crisis episodes have yielded two major lessons:

- It is of utmost importance to prevent the economy from sliding into paralysis and to avoid setting deflationary mechanisms in motion. The policy response needs to be of the "shock-and-awe" type. Monetary and fiscal policies can be powerful in alleviating the impact of the crisis in the short and medium run; however, as shown by the Japanese example, there cannot be a sustained recovery as long as banks are paralyzed and unable to extend credit (box 8.7).
- Partial injections of capital into the banking sector are of limited effectiveness as long as assets of uncertain value remain on the banks' balance sheets. Creditors remain wary of the soundness of the bank, which in turn leads it to err on the side of caution and restrict credit. A comprehensive cleaning up of balance sheets, and transparency as to their content and resilience to stress scenarios, are preconditions for credit revival.

Box 8.7 A Tale of Two Banking Crises: Sweden and Japan

While the crisis that erupted in 2007 was the first global crisis of this sort since the 1930s, it was by no means the first banking crisis in

modern times. On the contrary, there has been extensive international experience with such crises in developed and developing countries (Laeven and Valencia, 2008). Two examples were studied especially closely, those of Sweden in the early 1990s and Japan from the early 1990s to the mid 2000s. In both cases the country suffered from a severe banking crisis resulting in massive losses and the insolvency of a large part of the banking sector.

Measures introduced by the governments were broadly similar: In a first phase, liquidity was extended to ailing banks, a blanket guarantee of deposits was introduced to avoid panic; banks were thereafter nationalized, recapitalized through the injection of public funds, restructured, merged, and eventually privatized; and nonperforming assets were transferred to public asset-management companies in charge of selling them. However, the timing was very different: The Swedish government acted swiftly and decisively to ward off the crisis and adopted a hands-on approach to bank rescue and restructuring, while several years passed until the reality of the crisis was recognized in Japan, and even more years passed before the problem was addressed. Three years into the crisis, 50% of the losses had been recognized in the accounts of the Swedish banks, against 10% in Japan (figure B8.7.1).

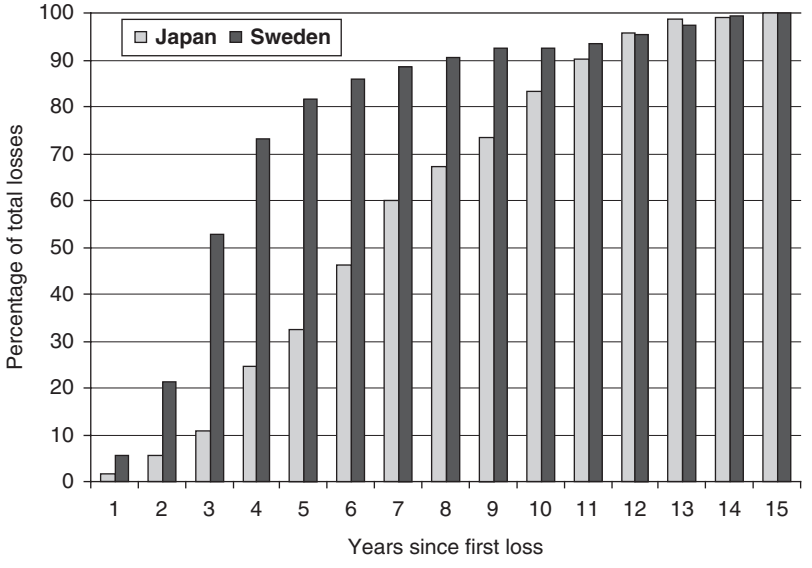


Figure B8.7.1 Cumulative bank write-downs: Sweden and Japan.
Source: Hoshi and Kashyap (2008), Bank of Sweden, and Bruegel calculations.

As a consequence, the crisis lasted longer and was significantly more costly in budgetary terms in Japan (table B8.7.1). The reason why

the outcome was so different is that delaying restructuring does not add to the chances of spontaneous recovery, rather it generally leads to further losses. For instance, credit restrictions raise the failure rate of enterprises, thus changing healthy loans into nonperforming ones. Then more capital needs to be injected into banks, creating more public debt while GDP—the denominator of the debt ratio—tends to stagnate. Japan thus seemingly managed to limit the short-term economic impact of the crisis as compared to Sweden, but at the price of a slow-growing GDP in subsequent years: 1.1% yearly growth on average between 1997 and 2006 for Japan, versus 1.7% and 2.6% over 1987–96 and 1997–2006 respectively for Sweden (source: OECD, Economic Outlook 85, June 2009).

Table B8.7.1
Cost of the banking crisis: Sweden and Japan

	Sweden	Japan
Start	1991	1997[a]
End[b]	1996	2005
Length (years)	≤ 5	≥ 8
Cost of bank recapitalisation (% of GDP)		
• Gross	1.9	6.7
• Net	1.5	6.6
Gross public debt ratio increase (start to end, percentage points)	39	76
Output loss (first three years, cumulative, in % of trend GDP)[c]	31	18

[a]The Japanese banking crisis began to develop following the stock-market crash in 1990 and the decline in real-estate prices, but the onset of the banking crisis is generally considered to be the failures of Sanyo Securities and Yamaichi Securities in November 1997 and resulting disturbances on the interbank market. It is only at that time that the extent of the damage began to be recognized.
[b]Date of removal of the blanket deposit guarantee (this tends to overestimate the duration of the crisis).
[c]Sum of the differences between trend and actual GDP over three years, divided by trend GDP.
Source: Laeven and Valencia (2008), OECD data, and authors' calculations.

As a consequence, government should intervene both on the liability side of banks' balance sheets through capital injections, refinancing, and bank debt guarantees, and on the asset side by buying assets or guaranteeing their value. All these instruments are detailed in box 8.8.

Box 8.8 A Primer on Bank Losses and Rescue

To understand the impact of bank losses and the options for government rescue, it is best to start from a very simple example. Assume the balance sheet of a bank prior to the crisis looks as in table B8.8.1:

Table B8.8.1
A bank balance sheet before the crisis

Assets		Liabilities	
Toxic assets	20	Equity	10
Other financial assets	20	Debt	50
Loans	50	Deposits	40
Cash	10		
Total	100	Total	100

On the asset side the bank holds cash, loans, and standard financial assets as well as *toxic assets** supposedly worth 20. The term "toxic assets" refers to assets whose market value is highly uncertain—although this may not be duly recognized in the absence of a crisis (for example, mortgage-backed securities whose yield depends on the stream of interests and repayments on mortgage loans to subprime creditors).

On the liability side it receives deposits from customers and issues debt. In this simplified example the difference between the market value of its assets and that of its liabilities is its equity, that is, the value of the bank's shares. It is assumed that assets and nonequity liabilities are accounted at market value.

Suppose now that the toxic assets held by the bank lose half of their value. The total assets of the bank are now worth 90 instead of 100 but liabilities to creditors and depositors have not diminished. This implies a loss of 10 on its profit-and-loss account and therefore a write-down on its capital that brings its equity to zero (see table B8.8.2). As a consequence it is bankrupt. It can repay its creditors and depositors by selling off its remaining assets (assuming they can be sold at their book value) but cannot remain in business.

Table B8.8.2
The bank incurs losses on "toxic" assets

Assets		Liabilities	
Toxic assets	**10**	Equity	**0**
Other financial assets	20	Debt	50
Loans	50	Deposits	40
Cash	10		
Total	90	Total	90

The bank can however refuse to recognize the extent of its losses and mark down its toxic assets at 15 instead of 10. This has two consequences:

- First, it is vulnerable to creditors' suspicion: customers may withdraw their money because they fear an outright default (this is what happened in 2008 to Northern Rock, the UK bank) and other banks may refuse to renew credit (this is what happened on the interbank market starting in August 2007). It is therefore likely to call on, and depend on, central bank credit as a substitute for private credit.
- Second, it is *undercapitalized**, because the loss of 5 that it has recognized on its assets implies a corresponding write-down on its equity. As a consequence the bank needs to raise capital or to reduce both its assets and nonequity liabilities to a level consistent with its remaining capital. This results in a nonrenewal of existing loans to clients and in a reduction of the volume of new loans.

"*Zombie banks**" of this sort are a dangerous species. First, they may at any time fail to meet their obligations and trigger a chain of defaults and therefore make the entire financial system more fragile. Second, they are inclined to ration credit and therefore impose costs on the nonfinancial sector. This is why swift government intervention is necessary to force banks to recognize their losses and operate a triage between the profitable, the viable, and the bankrupt.

Governments can intervene either through the liability- or through the asset side of the balance sheet. In the first case the most straightforward way to proceed is to nationalize the bank at no cost (since the value of its equity is zero) and inject new capital in the form of equity. In the absence of outright nationalization the government can inject capital through other channels such as *preferred stocks** or *preferred shares**.[a] Assuming the government injects both equity and preferred stock, the balance sheet now looks as follows (table B8.8.3):

Table B8.8.3
The bank is recapitalized by government

Assets		Liabilities	
Toxic assets	10	Equity	**5**
Other financial assets	20	Preferred stock	**5**
Loans	50	Debt	50
Cash	**20**	Deposits	40
Total	100	Total	100

Another way to proceed, if the government does not want to nationalize banks, is to purchase toxic assets at an inflated price (table B8.8.4).

For example, toxic assets can be isolated by setting up a *bad bank**.[b] This is another way to inject cash into the bank, but with very different distributional consequences. Instead of buying up a bank at zero cost (and possibly making a profit on its resale) the government buys toxic assets above market value and therefore makes a sure loss. The value of private shareholders' equity is thus indirectly subsidized, whereas they would be wiped out in the case of nationalization. These distributional consequences stand as political-economy arguments against setting up bad banks, even though this may be an effective solution to deal with toxic assets.

Table B8.8.4
Toxic assets are bought by government above market value

Assets		Liabilities	
Toxic assets	**0**	Equity	10
Other financial assets	20	Debt	50
Loans	50	Deposits	40
Cash	**30**		
Total	100	Total	100

[a]Preferred stocks are stocks which deliver a higher yield but which carry no voting rights. In case of bankruptcy, preferred stockholders are paid before stockholders and after bondholders.
[b]A "bad bank" is a temporary, public-funded financial structure designed to manage a set of assets taken out of ailing banks in order for the latter to be able to restart exposure to new risks through lending and to qualify as "good banks."

All types of intervention have been used to varying degrees during the crisis. A radical combination used by Sweden in the 1990s was to nationalize, remove toxic assets from banks' balance sheets, sell the banks back to the private sector, and use the proceeds to compensate for losses suffered on toxic assets. The ultimate net fiscal cost of the Swedish rescue plan was small: 1.5% of GDP as compared to 6.6% of GDP in Japan (box 8.7). However, only four banks were concerned at that time. Generalizing such a scheme to many banks in many countries was deemed impossible, notably when taking political constraints into account. Furthermore, in Europe, nationalizing banks with large cross-border activities would have required a level of coordination which could not be attained in the heat of battle and given the subsequent need to decide on how the fiscal burden would be shared.

Rather, bank recapitalization plans were carried out on a country-by-country basis, with striking differences of degree and procedure (table 8.2). In Europe, national initiatives to support banks were subject to speeded-up competition policy procedures for state aid, which resulted in a certain degree

Table 8.2
Bank rescue measures implemented in 2008–09 in selected countries (% of GDP)

	France	Germany	Ireland	The Netherlands	UK	US
Broadening of deposit insurance		Y	Y		Y	Y
Capital injections (effective)	1.2	2.0	6.5	6.8	2.6	2.1
Debt guarantees (effective)	5.5	7.2	164.7	7.7	11.3	2.4
Asset relief (effective)	0.2	1.4	0	5.5	14.7	0.3
Nationalizations		Y	Y	Y	Y	Y

Data cover the September 2008–August 2009 period. A blank cell means that the measure was not part of the rescue package. Y means it was part of it. 0 means that the measure is part of the rescue package but that there was no outlay during the period covered. Figures represent outlays and are given in % of GDP.
Sources: Pisani-Ferry and Sapir (2010) for the EU countries, on the basis of European Commission data, and Panetta et al. (2009).

of consistency across countries, but no discipline of this sort was implemented at global level.

A first reason for differences is the divergence in initial situations. In Spain, banks had been prevented by the bank supervisor from buying mortgage-backed securities and forced to build-up strong capital buffers during the housing market boom. They were affected by the collapse of housing prices but not by the subprime crisis. In Germany, regional banks had a weaker capital base and had invested heavily in structured assets. In the UK, mortgage refinancing by short-term borrowing on financial markets had been a flourishing business model that was destroyed by the crisis. In emerging economies, banks were scarcely exposed to structured assets and were only hit by the collapse of world trade and, in deficit countries, by the sudden stop of capital inflows.

A second reason lies in the structure of the banking system: In continental Europe, commercial banks with strong deposit bases were dominant, while investment banks led in the US.

A third reason has to do with political-economy constraints, which played a major role in determining the nature of the responses, both at a national and at an international level:

- At a national level, there is a trade-off between efficiency and equity. By providing generous recapitalization with little constraint in terms of governance, or by purchasing toxic assets at an inflated price, governments could quickly restore bank solvency and encourage private investors to invest in and lend to banks again. But the cost would then be borne by the taxpayer, while it should primarily be borne by the bank's shareholders, who had reaped generous revenues in the years before the crisis and had accepted the accompanying risk. Alternatively, refraining

from rescuing banks, or imposing a large cost on shareholders or, worse, on employees, would preserve the taxpayer in the short run but might fail to fix the problem, thereby inflating the final cost to the taxpayer. The US case illustrates this discussion. The Bush, then Obama administrations had a hard time convincing Congress to use taxpayers' money to support banks in the midst of a recession, and had to limit themselves to the initial 700 billion dollars allocation. As a consequence, they did not aim to maximize the return on public cash injections but to maximize the effect of injecting a given amount of public cash.

- At an international level, the absence of *ex ante* arrangements on sharing the fiscal cost of bank rescues makes it even more difficult to design them. As Charles Goodhart (2009, p. 16) said, "cross-border banks are international in life, but national in death." In such a context, tight international coordination of the supervision of cross-border institutions and of bank resolution regimes is called for (see section 8.3).

b) Unconventional monetary policy

In chapter 4 we explained how the Taylor rule could be used to provide a rough benchmark for setting the short-term interest rate. A standard formula relates this rate to the "equilibrium" real interest rate, the inflation rate, and the output gap.

Application of this benchmark rule would have resulted in significantly negative *nominal* interest rates in 2009 in the US, the euro area, and Japan. As already stated in section 8.1, monetary policy during the crisis thus encountered the *zero-bound problem**: While the Taylor rule would have recommended a negative interest rate, this is not possible to achieve, because depositors are not prepared to pay for keeping deposits with the banks (they would rather buy safes and keep cash at home).

In the 1990s, the Japanese experience with deflation and the liquidity trap prompted fresh thinking on the options still available when the interest rate cannot be lowered any further. The Fed especially studied this episode extensively and reached the conclusion that monetary policy could still be used and be effective. This was the origin of what became known as the *zero interest-rate policy* or *ZIRP**.

Another reason why unconventional methods are called for in a financial crisis is that the traditional transmission of policy rates to lending rates is hampered by the dysfunctional state of money markets. This happens at two levels: First, the interbank rate (the rate at which banks lend liquidity to each other) diverges from the central bank's policy rate because banks fearing counterparty default price risk accordingly; second, the spread between the commercial banks' lending rate and the interbank rate increases both because of higher risk premiums and because banks seek to increase their profits. Both phenomena were apparent in 2007–09 as illustrated by figure 8.5 for the UK: Prior to summer 2007 there was barely any difference between

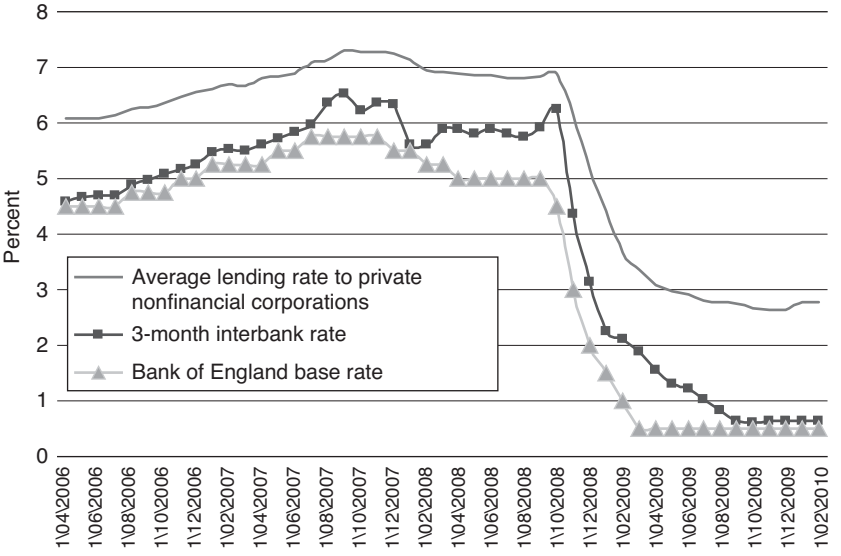


Figure 8.5 Three-month interest rates in the UK, 2006–10 (annualized, in %). Source: Bank of England.

the Bank of England policy rate and the interbank rate, but the spread widened in 2007–08 and reached 175 basis points in autumn 2008 (it only returned to negligible levels in summer 2009); simultaneously, the spread between the interbank rate and the banks' lending rate widened from 100 basis points in summer 2007 to 200 in summer 2009. The net result was that only about four-fifths of the 525 basis points policy rate cut was passed on to nonfinancial agents. Furthermore, quantitative restrictions were widely reported. So recourse to unconventional methods may be needed even before the policy rate hits the zero bound.

To understand ZIRP it is best to start with a simple thought experiment. Imagine that the central bank prints vast amounts of banknotes and drops them above big cities from helicopters. Surely, individuals receiving banknotes from heaven would feel suddenly richer and would spend at least part of this money (especially if they have heard about monetarism and fear that relying on the printing press will in the end result in inflation). Demand would pick up and inflation would follow later on with the consequence that the short-term real interest rate would decrease, leading to a further increase in demand.

What this thought experiment demonstrates is that the central bank's exclusive power to create money remains effective whatever the interest rate level and the state of money markets. Despite the fact that it does not provide the actual means to conduct monetary policy, it gives indications about what it can be. Surely, there must be more practicable ways to channel money to private agents than dropping banknotes from helicopters.

Policy thinking about unconventional policies was still fragmentary when the crisis hit (there had been debates and reflections about the Japanese experience but no systematic doctrine had been formulated, let alone a generally accepted definition of unconventional policy). Several approaches and partially overlapping concepts were therefore put forward in 2007–08 (Bernanke, 2009, and King, 2009 provide practitioners' rationalizations. Meier, 2009, gives a systematic account of the evidence).

The *large-scale provision of liquidity to financial institutions*, beyond the scale of normal operation of the discount window, is arguably more an adaptation of standard central-bank practice than a genuinely unconventional policy. Starting in summer 2007, all central banks extended wholesale liquidity to domestic financial agents. At an international level, liquidity provision also involved swap agreements between central banks, such as those entered into by the Fed with partner central banks in developed and emerging countries (box 8.9). Such agreements served a useful purpose in supplying banking systems with US dollars, while highlighting the lack of international coordination of last-resort liquidity provision (Obstfeld, 2009), an issue we will address in section 8.3.

The reason why, although truly exceptional, such initiatives do not fundamentally depart from standard monetary policy, is that they essentially aim at substituting the interbank market when it is clogged. Although they result in an increase in the size of the balance sheet of the central bank, they may leave constant the amount of money held by nonfinancial agents. In other words, the supply of base money (the central bank's balance sheet) has to increase because the ratio of money held by the public to base money (the *money multiplier*) has dropped due to reduced credit extended by commercial banks. This is what the Fed had failed to grasp during the Great Depression, thereby aggravating the crisis. Central banks this time fully offset the drop in the multiplier, without actually increasing money held by financial agents (von Hagen, 2009).

Box 8.9 International Swap Agreements

*Swap agreements** when entered into by major central banks enable partner central banks to provide commercial banks and other financial market participants with liquidity in foreign currency when they cannot obtain it on the market anymore.

A currency swap is a contract between two parties to exchange an asset in one currency for an asset of equal value denominated in another currency. When the Fed enters into a swap agreement with the ECB it supplies the latter with US dollars and takes an equivalent amount of euros in exchange. Swaps are entered into for a time-bound period.

In autumn 2008 existing US Federal Reserve swap agreements with the ECB, the Swiss National Bank, the Bank of England, and the Bank

of Japan were adjusted to unlimited amounts and new swap lines were extended to the central banks of Brazil, Korea, Mexico, New Zealand, and Singapore. The Fed's intention was to make sure that financial market participants operating in dollars on non-US markets could access dollar liquidity in spite of the clogging of interbank markets. The amount drawn by partner central banks exceeded 500 billion dollars in winter 2008–09 (figure B8.9.1).

In Europe the Swedish central bank entered into similar agreements with partners in Iceland, and in central and eastern Europe (Latvia and Estonia). The European Central Bank (ECB) remained more cautious: It established swap lines with central banks in Denmark, Sweden, and Switzerland but not with countries of central and Eastern Europe.

In May 2010, central banks had to revive swap agreements to help European banks access dollar liquidity in the wake of the Greek crisis.

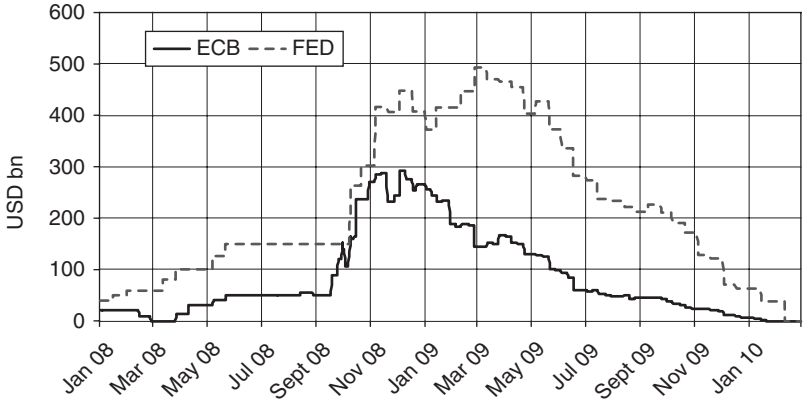


Figure B8.9.1 US dollars provided through swap lines by the ECB and the Fed, in US billions, 2008–10.
Source: ECB Monthly Bulletin, July 2009. Data available free of charge via the ECB's homepage.

According to Meier (2009), genuinely unconventional policies involve two types of actions:

- Announcements and/or refinancing operations designed to affect the yield curve at longer-than-usual horizons;
- Outright asset purchases, generally known as quantitative easing or credit easing, to reduce the spread between interbank and lending rates.

Central banks normally only target the short end of the yield curve, leaving the determination of longer-term interest rates to market mechanisms. In a

situation of near-deflation, however, expectations of positive interest rates and very low or negative inflation may combine to fuel a deflationary spiral. For this reason, central banks can commit to keep policy rates low for an extended period and enter into refinancing operations with extend maturity, possibly at a fixed rate and with unlimited amounts, thereby imposing a ceiling on interest rates at the corresponding horizon. This may imply committing, implicitly or explicitly, to higher inflation in the future, in order to lower expected real interest rates and encourage borrowing and investment.

These techniques, first suggested by Paul Krugman (1998a) and then-scholar (and later central bank governor in Cyprus) Athanasios Orphanides (2004) in the context of the Japanese crisis, have been used to varying extents by central banks, though none has gone as far as following Krugman's prescription and "committing to being irresponsible" (see chapter 4). For example, the US Federal Open Market Committee's statement of August 2009 included, as in previous months, the announcement that "the Committee continues to anticipate that economic conditions are likely to warrant exceptionally low levels of the federal funds rate for an extended period." The ECB used a different channel to lengthen the agents' horizon: In June 2009 it provided 12-month collateralized loans to the banks at a fixed 1% rate and for an unlimited amount (*ex post*, the banks' borrowing amounted to about 5% of GDP), but without committing to a repeat of this transaction.

Rather than aiming at affecting the overall yield curve through expectations of future rates, the central bank can directly affect yields on certain categories of assets through outright purchases. These can be either debt instruments issued by nonfinancial agents or government bonds. The rationale here can be to unfreeze clogged segments of financial markets, to help nonfinancial agents to get access to better and cheaper credit, and to affect long-term bond rates directly.

Meier (2009) provides a categorization of such operations, distinguishing between *qualitative easing** (sterilized interventions that do not involve an increase in the central bank's balance sheet) and *quantitative easing** (unsterilized interventions implying an increase in base money). Table 8.3 summarizes these various options and indicates what major central banks actually engaged in. The ECB stands apart for not contemplating quantitative easing (although its purchase of covered bonds may not have been sterilized entirely, the amounts potentially involved were a mere 0.6% of GDP). The Bank of England, the Bank of Japan, and the Federal Reserve engaged in significant quantitative easing with announced amounts of 8.6%, 5.2%, and 14.7% of GDP respectively. Finally, the Swiss National Bank stands out as the one that relied on unsterilized currency intervention.

Direct purchases of government bonds have a special status as they break the separation between monetary policy and fiscal policy and evoke debt monetization. They can be an effective tool when short-term interest rates are close to zero and longer-term rates well into positive territory: Government bond purchases can be effective in flattening the yield curve, which benefits

Table 8.3
Categories of unconventional monetary policy operations involving asset purchases

	No expansion of base money (qualitative easing)	Expansion of base money (quantitative easing)
Purchase of private assets (credit easing)	ECB	BoE, BoJ, Fed, SNB
Purchase of government bonds	ECB	BoE, BoJ, Fed
Purchase of foreign-currency assets (forex intervention)		SNB

Note: ECB, European Central Bank; BoE, Bank of England; BoJ, Bankof Japan; Fed, Federal Reserve; SNB, Swiss National Bank.
Source: Meier (2009), on the basis of announcements made by end-June 2009, and update by the authors.

all long-run borrowers, including corporations and foreign borrowers. Still, such a policy is normally taboo as it comes close to a direct financing of the government by the central bank—hence a monetization of the public debt as feared by Sargent and Wallace (see chapter 4). In the euro area, for example, the provision by central banks of credit facilities to governments or the direct purchase of government debt instruments are prohibited by Art 123 of the EU Treaty.[30] This taboo was broken in the US and the UK, as it was in Japan in the early 2000s. It was finally broken by the ECB too in May 2010, when it announced a program of government bond purchases after the crisis in Europe had morphed into a sovereign debt crisis affecting Greece, Portugal, and other southern members of the euro area. This program, however, was not launched for monetary policy purposes but to restore the functioning of certain national bond markets. It therefore amounted to qualitative easing (table 8.3).

Overall, liquidity provision and unconventional policies resulted in an unprecedented increase in the size of the central banks' balance sheets (figure 8.6). In spring 2009, assets held by the Federal Reserve and the Bank of England were more than twice as high as in spring 2007, and they were about 50% higher for the ECB.

Unconventional policies are necessary in exceptional circumstances, but they are not without risks:

- The direct inflation risk is less significant than often argued. The expansion of base money does not in itself create an inflation risk if it is undertaken in response to a reduction in the money multiplier. It can be

30. This is somewhat hypocritical since the eurosystem does purchase European government bonds for investment purposes (on the secondary market and in limited amounts), and since there is little economic difference between an outright purchase of a government bond and a liquidity tender with the same bond used as collateral, which can be rolled over as many times as needed.

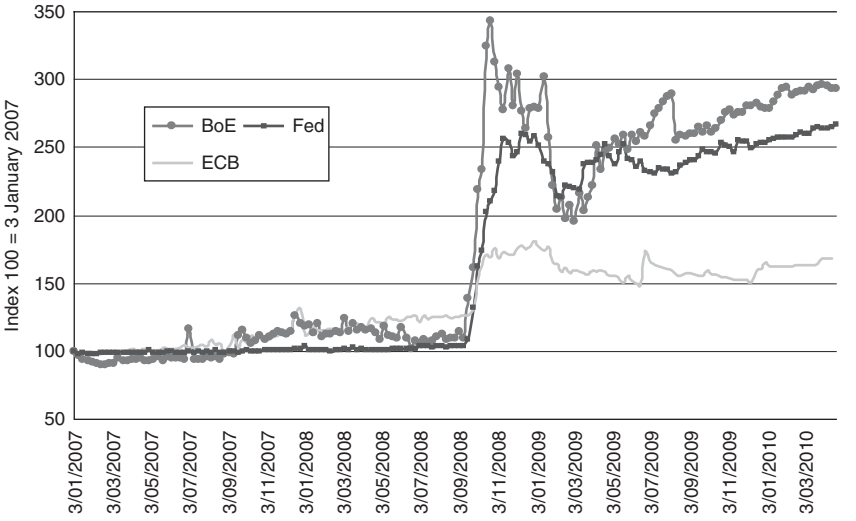


Figure 8.6 Total assets of selected central banks, 2007–09.
Source: Central banks.

reversed easily in response to a revival of direct interbank lending. Exceptional liquidity provision does not necessarily imply a more expansionary monetary policy.

- Through liquidity provision, the composition of central banks' assets has been massively skewed toward riskier assets. In principle, central banks apply an appropriate *haircut** (discount) to the collateral they take in order to account for the risk. Furthermore, the collateral remains the property of the banks and only serves as a guarantee for the central bank's loan. However in times of crisis the frontier between liquidity provision and subsidization is a thin one, and—if, for example, the risk is not adequately priced or if the market for the assets taken as collateral is paralyzed—the central bank can de facto become a *quasi-fiscal agent*, in effect blurring the distinction between monetary and budgetary policies. In particular, a loss on the assets bought by central banks could necessitate an intervention of the treasury to recapitalize it, thus endangering its independence.
- Commitments that affect the yield curve beyond the usual very-short-term horizon, or assets purchases that have the same goal, may involve an inflation risk. For the central bank, committing credibly to keeping interest rates at near-zero levels for an extended period amounts by definition to taking an inflation risk. This may be the price to pay for lowering *ex ante* real interest rates.
- By the same token, such policies break with the tradition that only the very short end of the yield curve is policy-determined and that the rest of it is market-determined. This may at a later stage make it difficult to

return to a policy of nonintervention in the formation of medium- and long-term interest rates.

c) Large-scale discretionary fiscal stimulus

As developed in chapter 3, prior to the crisis the effectiveness of fiscal policy was the subject of fierce debate. In the EU, conventional wisdom was that counter-cyclical fiscal policy was useful but should only rely on automatic stabilizers. Due to implementation delays and/or political cycles, discretionary fiscal policy was not considered an effective stabilization instrument. On each of the three criteria of flexibility, speed of action, and reversibility, it was outperformed by monetary policy. Even within the euro area where monetary policy was no longer available to respond to country-specific shocks, automatic stabilizers were considered large enough to stabilize country-specific shocks, provided public accounts were kept "close to balance or in surplus" in the medium run and were allowed to temporarily exceed the 3% deficit threshold in case of exceptional circumstances.[31] This explains why, when the crisis hit, many were not at ease with the very principle of a fiscal stimulus. Prominent policymakers such as Jean-Claude Juncker, the President of the Eurogroup, kept insisting that "you cannot fight debt with new debt and deficits with new deficits."[32]

Yet, unlike monetary policy, government demand for goods *directly* affects spending, thereby complementing rather than stimulating private demand. It can therefore be especially effective in situations when monetary policy effectiveness is hampered by a series of obstacles. Furthermore, in time of deep recession, many of the usual counter-arguments to discretionary fiscal policy do not apply:

- The magnitude of the drop in demand implies that there is virtually an excess supply of all goods and services in all countries and that inflation is decreasing. As a consequence, the supply curve can be considered flat (i.e., the supply can increase without any upward pressure on prices), the traditional crowding-out effects on investment and trade do not apply.
- As financial markets are dysfunctional, private agents are not able to borrow freely and engage in intertemporal optimization. More of them are liquidity constrained, as in the textbook Keynesian model.
- With unlimited credit supply by the central bank, there is no risk that public borrowing crowds out private borrowing.
- Cross-country externalities, whose signs are ambiguous under normal conditions, turn positive because spillovers through product markets dominate spillovers through capital markets.

31. See chapter 3 for a description of the European Stability and Growth Pact.
32. *Financial Times*, 4 April 2009.

In other words the macroeconomic conditions at end-2008 when the stimulus programs were launched were exactly those in which discretionary fiscal policy could be expected to be effective—provided that funds were disbursed swiftly enough. This was recognized by long-time advocates of fiscal policy ineffectiveness:

> Under normal circumstances, I would oppose this rise in the budget deficit and the higher level of government spending. When an economy is closer to full employment, government borrowing to finance budget deficits can crowd out private investment that would raise productivity and the standard of living. Budget deficits automatically increase government debt, requiring higher future taxes to pay the interest on that debt. The resulting higher tax rates distort economic incentives and thus weaken future economic performance. . . . Nevertheless, I support the use of fiscal stimulus in the US, because the current recession is much deeper than and different from previous downturns.
>
> Martin Feldstein (2009)

In Europe in November 2008, governments and the European Commission engineered an exceptional coordinated stimulus of about 1.5% of GDP. At about the same time the IMF advocated a 2% of GDP stimulus in all countries in a position to engage in such an action. In the US, the Obama administration introduced a two-year package amounting to 787 billion dollars shortly after taking office in January 2009. China also announced a massive stimulus program. As a whole, the Horton et al. (2009) estimate that G20 countries provided a discretionary impulse of 2.0% of GDP in 2009.

A distinctive feature of the coordinated stimulus of 2009 was that it was by no means restricted to the advanced countries. On the contrary, it involved significant participation by major emerging countries, whose reliance on discretionary stimulus generally exceeded that of advanced countries (figure 8.7). This was in part due to the lesser importance of automatic stabilizers in countries where social insurance systems are less developed and the state overall represents a lower share of GDP. But beyond this composition effect, a major policy change, largely engineered by the G20, was that emerging and developing countries took part in global demand management at world level, either by their own means (for example in the case of China) or thanks to loans provided by multilateral development banks (for example in the case of Indonesia).

As regards the composition of the stimulus, many countries put emphasis on public investment (both infrastructure building and incentives to private investment, especially "green" investments). The idea was to maximize the Keynesian multiplier and increase public assets simultaneously with public debts (so that net public debt would not rise too much). However, there are often delays in the implementation of public investment plans. For instance, the US Congressional Budget Office calculated in June 2009 that US expenditures on infrastructure building within the American Recovery and

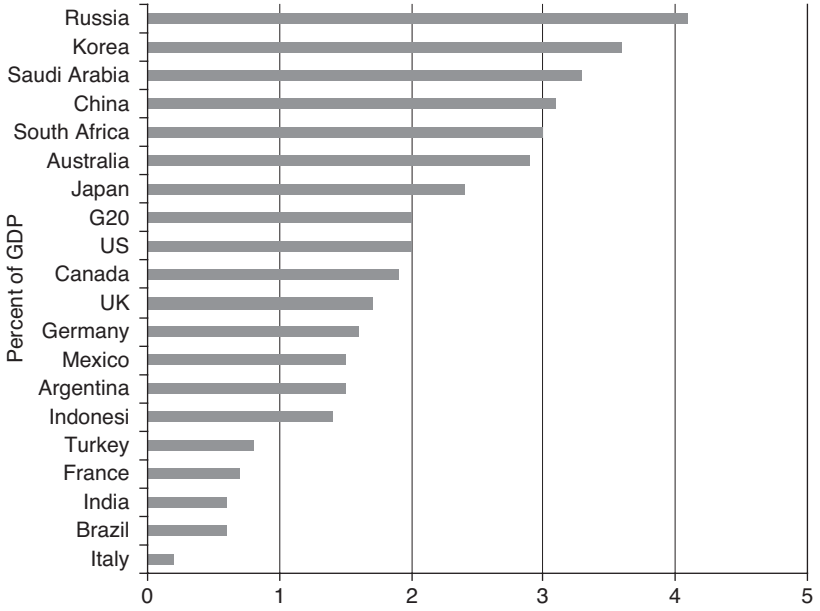


Figure 8.7 Size of 2009 fiscal stimulus plans in G20 countries.
Source: IMF (Horton et al. 2009), Bruegel calculations.

Reinvestment Act passed in February 2009 would peak in 2010 and 2011.[33] Conversely, some countries such as the UK relied primarily on tax cuts, which are very rapid to implement but may not translate into higher demand if private agents choose to save or, in the case of the UK VAT cut, may result in limited pass-through on prices if competition conditions allow suppliers of goods and services to retain the benefit of the cut. Additionally, tax cuts may be politically difficult to reverse.

Not all governments and central banks were able to turn expansionary. Central and eastern European countries were hit by "sudden stops" of capital inflows that forced them to reduce domestic demand through fiscal retrenchment and a tight monetary policy, even though they were supported by IMF and EU loans.[34] Developing countries that had resisted the crisis but which could no longer borrow from international markets were encouraged to carry out counter-cyclical fiscal policies with the financial support of multilateral and bilateral development banks such as the World Bank and

33. China did not experience such a delay because many infrastructure projects had been halted before the crisis when the government was aiming to slow down the economy. These were ready to be implemented when the crisis arose.
34. The countries on a fixed-peg regime also chose to defend their exchange rates through high interest rates.

the Asian Development Bank[35] and some of them were awarded the newly created "flexible credit line" by the IMF, a contingent financing facility (see chapter 5). IMF-subsidized loans to low-income countries were doubled. This was the first time official financing was extended as a support to counter-cyclical policies. More broadly, for the first time, the IMF vocally advocated large fiscal stimulus and bank rescue plans.[36]

Reliance on large-scale stimulus, coming on top of the cost of large-scale bank bailouts and of the recession-induced fall in tax receipts, led to a sharp increase in public debt ratios (figure 8.8).

The Irish case is especially dramatic since the tripling of the gross-debt ratio was accompanied by public guarantees extended to banks worth 200% of GDP. This case is by no means exceptional, though. Again, experience from past crises is unequivocal. The Japanese gross public-debt ratio rose from 64% in 1991 to 175% in 2005 as a consequence of the financial crisis and a series of stimulus packages (see also box 8.7). More generally, financial crises have been found to have large-scale consequences on public debt (Reinhart and Rogoff, 2009a,b).

At the end of 2008 bond markets started to discriminate more between euro area sovereign borrowers, while rating agencies downgraded several of them (see figure B3.1.1 of chapter 3). Initially these moves in part reflected an across-the-board repricing of risk after the collapse of Lehman Brothers, and in part a general lack of liquidity which favored the most liquid debts,

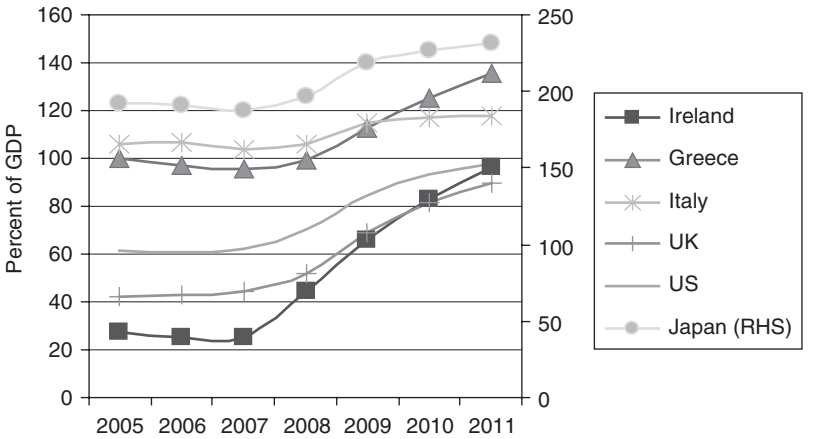


Figure 8.8 Gross public debt ratios in selected countries (% of GDP), 2005–11. Sources: IMF World Economic Outlook, April 2010; European Commission autumn forecasts, October 2009.

35. The International Financing Corporation (a branch of the World Bank group) and the European Bank for Reconstruction and Development also provided direct support to private sectors in developing countries, e.g., by providing fresh capital to banks.
36. See, for instance, Blanchard et al. (2008).

such as Germany and France. Increasingly, however, genuine concerns over public-finance sustainability in specific countries became the main factor of differentiation and this became evident in early 2010 when spreads began to widen again. Such concern was especially worrying for members of the euro area such as Greece, and also Ireland and Portugal, which have lost the ability to monetize public debts as well as policy space for macroeconomic support (see chapter 5). In countries with independent currencies such as the US, the concern over debt sustainability rapidly translated into a concern over inflation in the medium run. These two opposite cases can be seen as potential illustrations of the "game of chicken" depicted in the "unpleasant monetary arithmetics" model of Sargent and Wallace (see chapter 4, box 4.11).

To prevent such unpleasant outcomes while providing the required budgetary support in the short run, two-handed policies were called for: They needed at the same time to sustain significant spending programs as long as the recovery was not solidly under way, and to ensure sustainability in the medium run through credible commitments to reverse course in the medium run and bring public finances back to balance. In fact, analysis of the requirements of fiscal-policy effectiveness presented in chapter 3 suggests that the more sustainability is guaranteed for the medium run, the more stimulus packages are effective in the short run. So there is no contradiction but rather complementarity between providing Keynesian support and adhering to fiscal discipline. Still, such discipline is difficult to define in a credible way in the midst of a crisis. We will come back to this challenge in the next section.

8.2.2 The aftermath

a) Exit strategies

Exceptional challenges require exceptional responses, with the risk of building up distortions and disequilibria calling for later adjustment. There are many examples: Consolidation in the banking sector may hamper competition, inflated central-bank balance sheets may undermine confidence in price stability, stimulus packages and guarantees extended to the private sector may lead to an unsustainable build-up of public debt. Such concerns are of second order in the midst of the crisis, but they gain prominence along the recovery path.

The *exit strategy** issue raises difficult and related questions as to when, to what extent, at what pace, and in what order to unwind the unorthodox macroeconomic and financial policies undertaken in response to the crisis.

- Rather than being time-contingent, exit strategies should as much as possible be state-contingent. Public participation in the capital of banks and other support measures need to be maintained as long as banks remain too fragile to elicit confidence in capital markets. The experience of past crises shows that budgetary and monetary support should also be sustained as long as the recovery has not gained sufficient

autonomous traction. Earlier mistakes are telling: In 1936, the Fed severely tightened monetary policy by raising reserve requirements in order to check the expansion of credit. This killed the recovery that had started in 1933 and led to the 1937–38 economic contraction.[37] Japan also experienced a failed exit in 1997 (see below). Finding the right timing for policy reversal may, however, be tricky. State-contingent strategies may lose effectiveness if sustainability concerns take precedence, as confirmed in spring 2010 by the euro area crisis. It is also difficult to determine what the desirable course of policy action is when the impact of the crisis on potential output is uncertain (see the discussion below on the legacy of the crisis). If potential output has been lastingly dented by the crisis, then the output gap is smaller in absolute value. There is less need for demand stimulation and more need for supply-oriented reforms, without which inflationary pressures may build up sooner than expected.

- There is little debate over the need for fiscal policy to get back to normality as quickly as possible, which, beyond removing the stimulus, implies a large-scale retrenchment as public finances suffer from permanently lower revenues and a higher level of public debt. How to implement this retrenchment without endangering potential and actual growth is a major challenge for the medium term.
- As regards monetary policy, while wholesale liquidity support needs to be unwound, the very definition of the objectives and operational guidelines of policy is bound to be modified by the crisis. So there is certainly an exit, but not exactly to the status quo ante. This is even truer for the micro-interventions: Exit from government ownership requires that regulation be reformed and reinforced. Hence, a successful exit strategy is not a reversal to the *ex ante* policy framework.
- Finally, the *sequencing* question is the most daunting challenge. In principle it is advisable to start by removing the most distortionary components of the rescue packages, i.e., their micro-components, then remove the fiscal support (as it involves significant costs to public finances) and finally remove the monetary support. However political-economy considerations suggest the reverse order is more likely, because central banks will be keen on getting back on track; governments will be under pressure not to raise taxes and cut expenditures; and pressures to retain pervasive state intervention in the financial sector may remain strong in some countries. Priority for the fiscal exit also implies that monetary policy is bound to remain supportive for a long time, which invoves risks to financial stability.

Exiting also means ending the confusion of roles between monetary and fiscal policies. As already mentioned, in 2008–09 central banks inflated their balance sheets during the crisis and skewed their composition toward

37. For an account, see Friedman and Schwartz (1963). See also Kroszner (2009).

riskier assets. In the euro area, during the crisis, the ECB accepted government securities of lower quality as collateral, which put it in a trap: In spring 2010 it realized that returning to standard quality requirements would have risked excluding Greek bonds in the case of further downgrades by rating agencies, provoking a sell-off of these bonds and a default by Greece, but to keep governments bonds of lower quality on its balance sheet would have amounted to quasi-budgetary support. In March 2010 the ECB decided not to return to its standard quality requirements but to introduce graduated haircuts instead, thereby de facto opting for the second risk. In May, it started to buy Greek, Portuguese, and Spanish government bonds.

Finally, there are issues of international cooperation. Countries may exit the recession at a different pace depending on their initial situation (in particular as regards the degree of leverage of domestic private agents), policy responses, and exposures to the global shock. This calls for differentiated exit strategies—except where the crisis response affects internationally integrated market segments. This exception obviously applies to intervention on international capital markets (such as central-bank purchase of bonds and other securities) but also to government support to sectors highly exposed to international competition, such as the car industry. Europe is a special case: For example, early fiscal adjustment in a country may hamper the recovery of its partners; conversely, lack of adjustment in a large country or group of countries may lead to higher interest rates and exert negative externalities (see box 2.14 of chapter 2). By the same token, restoring a level playing field in the banking sector will require ending public support in a coordinated manner.

b) The legacy

International experience with financial crises suggests that there is a high risk of permanent potential output reduction. At a first stage, the sudden fall in output translates into bankruptcies, a rise in unemployment as well as workers exiting the labor market, and lower capital expenditures translating into a lower capital stock. Depending on economic institutions and on policies implemented in the aftermath of the crisis, the shock may in turn result in a permanently lower employment rate and a permanently lower stock of capital and technologies. Post-crisis policies should aim to limit the extent of this permanent damage.

Potential output. Returning to pre-crisis GDP levels requires several years of growth: Three years on average for the industrialized countries that experienced major financial shocks in recent times—Finland (1991), Japan (1997), Korea (1997), and Sweden (1991) (figure 8.9). In these cases, the recovery was largely driven by productivity gains,[38] whereas employment

38. The increase in GDP per hour worked evidenced by figure 8.9 can be partly due to a composition effect, since low-productivity workers are the first victims of a recession.

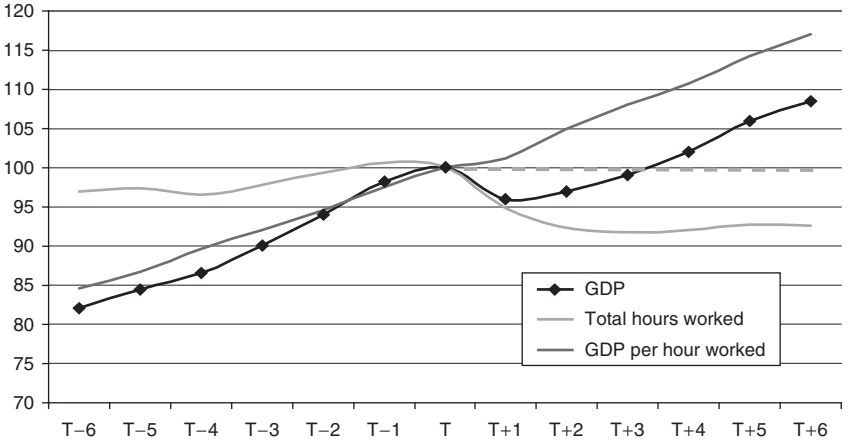


Figure 8.9 GDP profile through financial crises: Finland, Japan, Korea and Sweden. Average of four countries (100 = level in the first year of crisis).
Source: Pisani-Ferry and Van Pottelsberghe (2009) on the basis of national data.

lagged behind. In Finland, the unemployment rate of males jumped from 3.5% in 1990 to 7.9% in 1991. It peaked at 17.9% in 1993 and 1994, and was still 9% 10 years after the beginning of the crisis, according to the OECD. Hence, while GDP recovered relatively quickly, the crisis had a long-lasting effect on unemployment. This illustrates the risk of *hysteresis* of high-unemployment periods: In the process idle workers lose some of their skills or they cannot update them; those near the retirement age withdraw from the labor market; and the crisis first wipes out weaker industries, potentially aggravating unemployment of low-skilled workers.

The crisis will have lasting effects on the financial industry. Fewer actors and therefore less competition in the banking sector compounded with higher regulatory capital requirements will translate into higher borrowing costs, hence less investment and lower potential output. Risk-aversion may be durably higher, which may be helpful to prevent similar crises occurring, but detrimental to venture capital in innovative industries. As discussed in chapter 6, fluid capital markets are key to ensuring factor reallocation and risk-sharing in the innovation process; pre-crisis capital markets may have been too fluid, but a clogged financial system will not help either. If stability is achieved through tougher supervision and higher capital requirements, one result of the crisis will be higher capital costs, hence slower capital accumulation. This was an important dimension of the 2010 discussion on new capital adequacy requirements for banks ("Basel 3").

In some countries a shrinking financial sector could in itself reduce potential output. In Ireland, for instance, financial intermediation represented

10% of GDP in 2005.[39] Before production factors are reallocated to other sectors, a 20% reduction in value added in this sector would thus translate into a 2% fall in potential output. Additionally, the pace of potential output may have been overestimated during the pre-crisis boom, for instance due to overinvestment. Hence, GDP may not recover its pre-crisis level for many years.

The ability to reallocate labor and capital will be essential in order to limit the permanent consequences of the crisis. If successful, such reallocation could in principle *increase* potential output in the medium term (this is the "creative destruction" mechanism identified by Schumpeter, see chapter 6). Traditionally, Anglo-Saxon countries are more successful than those of continental Europe in reallocating labor across sectors and also geographically. The high unemployment in continental Europe in the 1980s and the 1990s has been ascribed to the interaction of adverse shocks and rigid labor market institutions (Blanchard and Wolfers, 2000). In these countries, further labor-market reforms will therefore be needed to bring unemployment rates back to their pre-crisis levels. The problem is that structural reforms are costly in the short run and may be politically difficult to implement in the aftermath of a crisis. Finally, large-scale reallocation of capital to new industries necessitates well-functioning financial markets and banks. It may be hampered by convalescent, more risk-adverse banks.

Public debt. Already burdened with inflated debts, governments will still face the cost of ageing, a cost that itself has been magnified by the crisis. The net cost of ageing for public finances is still valued at several percentage points of GDP.[40] Reforms of pension systems such as longer working periods are necessary.

This already existing challenge is compounded by the effects of the crisis. To start with, according to the International Monetary Fund (2009), the loss incurred by US and UK pension funds in 2008 amounts to 22 and 31% of GDP, respectively (excluding losses on toxic assets). Depending on how financial markets recover, there will be effects on public finances, both direct (because of unbalanced public pension funds or public bail-outs of private schemes) and indirect (through pressures for more generous pensions from the pay-as-you-go pillars to compensate for the reduction of funded pensions, and higher unemployment among older workers making pension reforms more difficult to engineer). Second, lower potential output makes it even more difficult to consolidate public finances. Let us suppose that industrialized

39. Source: EUKLEMS database, 2005 figure. The same year, financial intermediation represented 25% of GDP in Luxembourg, 8% in the UK and in The Netherlands, but only 2% in Finland.

40. Equivalent to an increase in the fiscal deficit of 2.9% of GDP from 2005 to 2050 in the US, 3.4% in the UK, 3.8% in Australia, 7.7% in Canada, and up to 13.4% in Korea (see International Monetary Fund, 2009).

countries have permanently lost 4% of potential GDP as a result of the crisis (see OECD, 2010). With government revenues amounting to around 35% of GDP (see chapter 7), this implies a permanent revenue loss equivalent to 1.4% of potential GDP, hence on average a structural deficit which is permanently higher by the same amount (it is significantly higher in some countries).

In plain English, the fiscal hurdle is now much higher. Lastly, and *in addition to the previous mechanisms*, lower growth speeds up debt accumulation by increasing the fiscal surplus required to stabilize the debt-to-GDP ratio (see chapter 3 for a description of pre-crisis debt dynamics).[41]

The functioning of the European monetary union. The crisis represents a particularly acute challenge for the EU. In the euro area, the Greek crisis that broke out in 2010 exposed the ambiguities and deficiencies of the EU treaties. In the absence of an institutionalized crisis-management mechanism, and given the ambiguous wording of the so-called "no bail-out" clause of the EU Treaty (see chapter 3), long negotiations were necessary to determine whether Greece could receive financial assistance from its EU partners, whether the IMF would participate in the assistance package, and what would be the terms of assistance. The absence of a crisis resolution mechanism was indicative of the incomplete character of the Economic and Monetary Union. At the same time, the economic crisis highlighted the extent to which protracted real exchange-rate differentials had developed within the euro area and put the weakest countries in severely uncompetitive positions, calling for a review process of countries' competitiveness as a complement to the Stability and Growth Pact.

Globalization. Finally, the post-crisis world will have to cope with possible "de-globalization" and rising protectionist tensions.

The sharp drop in trade in goods and services observed in late 2008 and early 2009 can be explained by the fall in world output, the shift of global demand away from capital goods (that happen to be traded more than consumption goods), the shortage of trade finance, and relative price effects.[42] Despite anecdotal evidence of rising tariff or nontariff barriers, genuine protectionism has not emerged as a response to the crisis (in contrast to the 1930s). Nevertheless, a failure of governments to curb unemployment could later on give rise to serious protectionist pressures that could significantly affect the globalization process at work during the 1990s and 2000s. Additionally, a sustained recovery of world growth will be very dependent on the ability of large, emerging countries to substitute for the US as a world growth engine.

41. Adverse debt dynamics can be mitigated to some extent by lower interest rates as a result of higher savings.

42. See Bénassy-Quéré et al. (2009).

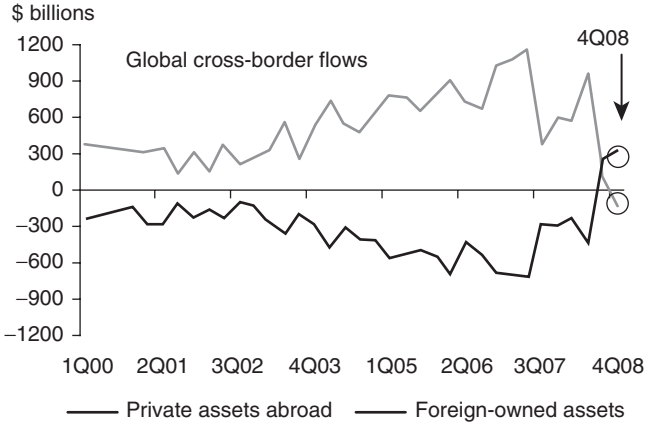


Figure 8.10 Financial de-globalization, 2000–08.
Source: Broda et al. (2009). ©VoxEU.org.
Note: Sum of variations of assets invested abroad and foreign-owned assets of the largest 30 countries by portfolio volume.

This means less outward-oriented growth models. For instance, China will need to lower its savings rate and develop its domestic market.[43]

On the financial side, large cross-border financial institutions will be less fashionable, since they are difficult to supervise and even more difficult to close when they go bankrupt. Small "host" countries will be more reluctant to host them, given the cost of a rescue in terms of their own GDP (the "too-big-to-save" syndrome) and large "home" countries will be keen on concentrating capital and liquidity in order to comply with the new regulatory framework. Also, the crisis has led to a dramatic reduction in cross-border investments (see figure 8.10). Developing countries have been especially hurt. As they started recovering sooner and at a faster pace than advanced economies, they subsequently received large capital inflows that sometimes contributed to an overheating of the economy. Stabilizing international capital flows along a North–South and South–South route rather than a South–North one will be a major challenge. For this purpose, emerging countries could diversify their foreign-currency holdings out of US Treasury bills and into "strategic" investments in energy, agriculture, manufacturing, and finance, using sovereign wealth funds as their instruments. This would be consistent with their climb up the quality ladder and the adjustment of

43. This means not only the reconstruction of social safety nets (to make it possible for households to save less), but also more incentives for domestic entrepreneurs to work for the domestic market (e.g., a halt to tax incentives for exports and a more robust legal framework for business domestically) and increased incentives for domestic banks to lend to domestic entrepreneurs rather than (as they have been doing again under the fiscal stimulus plan) to lend primarily to public enterprises and local governments.

their production structure. Such a transition would, however, require a large exchange-rate adjustment and may raise protectionist concerns in developed countries. Finally, absent "financial safety nets" protecting emerging market countries from international liquidity shocks, they will be willing to build-up even larger foreign-exchange reserves and/or restrict short-term capital flows.

8.2.3 Lessons

Before the crisis, policy discussion in advanced economies tended to favor rules over discretion, to downplay the role of fiscal policy as a way to stimulate aggregate demand, and to refrain from interfering with market signals. In sharp contrast, governments during the crisis held policy rules in abeyance, engineered massive bank rescues and industry-level support (such as in the car sector), launched fiscal stimulus plans, and took direct control of entire segments of the economy. As for central banks, they promptly played their role as lenders of last resort, brought refinancing rates to zero, and embarked upon previously untested unconventional monetary easing.

This course of action proved highly successful in comparison with the Great Depression—at least in the short run. The long-run consequences have not been yet tested.

The crisis has shown the importance of economic-policy rules going alongside escape clauses that allow policymakers to revert to discretion in exceptional times. However, it may also leave its imprint on the rules themselves. To provide but a few examples: The Stability and Growth Pact will not rein in government deficits in Europe without better enforcement, stronger ownership, and additional rules to cope with the higher cost of ageing, intra-Eurozone imbalances (such as cumulated drifts in price competitiveness) will have to be monitored more closely, and the crisis has re-opened the debate on asset prices and monetary policy rules. The shape of the post-crisis world is the subject of the next section.

8.3 In Search of a New Regime

We have analyzed in section 8.1 why the crisis can be attributed to microeconomic and macroeconomic factors, as well as to a lack of resilience of the system as a whole. We have described in section 8.2 the immediate reaction of policymakers to its outbreak. We now turn to the global reform agenda, that is, to the changes that are needed to reduce the risk of similar crises in the future.

This agenda was first defined in a series of meetings of the G20 leaders.[44] In an unprecedented exercise of international coordination they met in

44. For more on the G20, see chapter 2.

Washington, London, and Pittsburgh (Pennsylvania) in 2008 and 2009.[45] While a large part of these meetings was devoted to crisis management and financial regulatory reform, the reform of international financial institutions and the creation of a framework for international coordination ranked high on the agenda in order to "lay the foundation for reform to help to ensure that a global crisis, such as this one, does not happen again."[46] Specialized institutions such as the IMF, the Financial Stability Board (FSB), or the Bank for International Settlements (BIS) actively participated in the search for a better international financial and macroeconomic system.

This section presents the steps taken during these hectic times and discusses those that are still a matter for debate. We begin with the financial system and continue with the macroeconomic and monetary policy regime, before highlighting a number of policy issues that initially at least received relatively less attention from policymakers.

8.3.1 The financial system

As explained in section 8.1, the crisis revealed five major failures of the financial system:

- Excess confidence in self-regulation and a lack of regulation of some financial institutions, such as the special investment vehicles;
- Ill-designed regulation leading to excessive risk-taking and procyclical behavior;
- Weaknesses in the crisis resolution regimes (e.g., a lack of preparation for bank failures), and the nonexistence of serious schemes for addressing the failure of banks with significant cross-border activities;
- Gaps in financial infrastructures (e.g., the lack of organized credit default-swap markets);
- The absence of a systemic view of financial vulnerabilities.

Solutions were suggested to tackle each of these various failures of the system at the national, regional, and international levels.

a) The regulatory architecture

> We all agree that at the heart of the modern enterprise challenge is minimizing regulatory concerns. . . . The better, and in my opinion the correct, modern model of regulation—the risk-based approach—is based on trust in the responsible company, the engaged employee and the educated consumer, leading government to focus its attention where it should: no inspection

45. The scope and intensity of international coordination contrasts starkly with the response given to the crisis in the 1930s. Then, the one major attempt to organize a common response, the London conference of 1933, ended in disagreement, not least because Franklin D. Roosevelt deliberately torpedoed it.
46. Declaration of the Washington summit, November 2008.

> without justification, no form-filling without justification, and no information requirements without justification, not just a light touch but a limited touch.
>
> UK Chancellor of the Exchequer Gordon Brown (2005)

> No longer can we allow Wall Street wrongdoers to slip through regulatory cracks. No longer can we allow special interests to put their thumbs on the economic scales.
>
> President-elect Barack Obama (2009)

Prior to the crisis, regulation was often regarded as a hindrance to business and international competitiveness, and great trust was put in *self-regulation**. Examples of such confidence can be found in the Basel II capital standards for banks, which partially relied on in-house assessment of risks (see box 8.2), in the loopholes of banking supervision (such as the shadow banking system and US mortgage originators), in the high leverage ratios investment banks were allowed to maintain in spite of capital requirements, or in the leniency of regulators vis-à-vis the credit-rating agencies, whose performance was not monitored and whose conflicts of interest were not addressed.[47] These are only examples.

One clear manifestation of self-regulation failure, noted by Persaud (2009), was that, prior to the crisis, the equity of banks that were either bailed out or went bankrupt during the crisis, such as Northern Rock, Bear Stearns, Fortis, and Lehman Brothers, exhibited higher price–earnings ratios than those of more resilient banks such as HSBC or JP Morgan-Chase. The risks taken by the former group of banks were apparently not priced in by financial markets.

Regulatory loopholes are probably the most straightforward root of the crisis. Banks escaped capital regulations by using off-balance-sheet special investment vehicles (SIVs) to buy assets while financing these investments through the issuance of short-term asset-backed securities (ABSs). However, the corresponding risk was not transferred, since banks extended guarantees to their SIVs, or even held asset-backed securities while transferring their loans to SIVs in order to reduce on-balance-sheet risk. In brief, this is as if banks themselves had bought ABSs without capital. When in the wake of the crisis the short-run funding dried up, ABSs (from then on called "toxic assets") had to be transferred back to banks' balance sheets, where capital requirements apply, leading to sudden undercapitalization of the banking sector and to the subsequent disruptions in financial markets.[48]

Having understood the responsibility of regulatory gaps in the crisis, policymakers soon declared their intention to regulate all significant financial actors and markets. Advocates of free markets objected that bureaucrats are ill-placed to know what is good for the market and that there was a risk that excess regulation would hinder innovation and growth. However, this

47. Credit-rating agencies were not subject to regulation in the EU and had only to be registered in the US.

48. See Acharya and Richardson (2009).

objection can be circumvented by: (i) Delegating supervision and oversight to independent, technically able specialized agencies, and (ii) retaining elements of self-regulation in the *standard-setting* process, while giving responsibility for *enforcement* to independent supervisors.[49]

Closing the gaps in regulation requires comprehensive reform of the financial regulatory architecture, such as that proposed in 2008 in India by a reform committee chaired by former IMF chief economist Raghuram Rajan, in the US by the Treasury Department (US Treasury, 2009), and in the EU by the De Larosière report, subsequently endorsed by the European Council (De Larosière, 2009). One way to proceed has been to address regulatory fragmentation, which favors *regulatory competition** both within national borders (e.g., in the US) and across countries (e.g., in the EU):

- In the US, as many as five institutions—the Federal Reserve, the Federal Deposit Insurance Corporation (FDIC), the Office of Thrift Supervision, the Office of the Controller of the Currency, and the Securities and Exchange Commission—were responsible for banking supervision. Fragmentation favored regulatory arbitrage, especially as agencies were competing with each other for business. This number was reduced to three (the Fed, the FDIC, and the new National Bank Supervisor) in the Obama reform proposals of 2009.
- In the EU, cross-border institutions were supervised by banking supervisors, market regulators, and insurance regulators of all 27 member states. Following the De Larosière report, it was decided to transform existing committees of supervisors into three EU-wide bodies with extensive responsibilities.

In the decades before the crisis, *international regulatory competition* had become widespread, resulting in less-effective regulation. Attracting financial business was an integral part of competitiveness policies in a country like the UK, where light-touch regulation was viewed as a way to create and attract jobs. This logic was pushed to the extreme in *regulatory havens**, or *non-cooperative jurisdictions**, i.e., countries or territories which operate a financial industry and do not enforce the standards produced by international regulatory agencies. In the aftermath of the crisis, G20 countries agreed to put pressure on the latter countries.[50]

49. Credit rating agencies (CRAs) are an example of such an approach. Following extensive consultation with the industry, the International Organization of Securities Commissions (IOSCO) has elaborated a code of conduct for CRAs to address conflict of interest between investor information and credit structuring advice, and to ensure transparency of rating performance. CRAs now have to register with the market regulator, and they will be de-registered if they do not comply with the IOSCO code of conduct.

50. Although the two concepts partially overlap, regulatory havens should not be confused with *tax haven*. The G20 also took fierce measures against the latter, but the rationale there was to repatriate tax bases at a time when tax receipts were experiencing a steep fall. Non-cooperative jurisdictions

Regulatory coordination also helps contain *regulatory (and supervisory) capture**. National authorities may be excessively lenient toward domestic institutions, either because they favor national champions or because of a blurring of political and business interests. The case of France, where political elites often push expansion of domestic banks in the name of the national interest; of Germany, where politicians sit on the boards of *Landesbanken*; or of Italy, where central bank governor Antonio Fazio was forced to resign in 2005 after it became clear that he had used financial stability arguments to protect the business interest of certain Italian banks, all show that such practices are a fact of life. International regulatory harmonization and cooperation helps limit their extent.[51]

There will always remain free riders. Contrary to rogue nuclear states, some of them have good excuses: Offshore finance is sometimes their only specialization. A way out of this problem is to blend cooperation and incentives, e.g., through internationally sponsored technical assistance to improve their regulatory capacities and to help them change their model, and of suasion, e.g., by making it more costly for international banks to do business in recalcitrant territories.

A related, though not identical, problem is the assignment of supervisory responsibility. Banks are generally supervised by the authorities of the country where they are headquartered, at least when they operate through branches and not legally independent subsidiaries. But financial stability is the responsibility of *host*-country authorities. It has therefore been suggested that responsibility for supervision should switch from home to host country. This would require banks to operate through subsidiaries rather than branches, with each subsidiary being regulated and supervised by the host country. An additional advantage of such an approach would be to help each host country engineer macro-prudential supervision in relation to its own credit cycle. The danger, however, is potential financial fragmentation, possibly reducing the scope for economies of scale and opening the way to financial protectionism.[52]

The reform of the regulatory architecture is by its nature an unfinished agenda and the tension between, on the one hand, worldwide market integration and, on the other hand, national sovereignty and different national preferences, is here to stay. Some like Dani Rodrik (2009) think that this tension is in fact an inherent contradiction, and that attempts at international regulatory coordination are bound to fail. At any rate, in a situation where public opinion is expressing increasing anger against those who provoked the crisis, there will be inevitable frictions between national politics and international initiatives.

---

include tax havens regulatory havens, and territories that do not enforce international rules against money-laundering and the financing of terrorism.

51. This illustrates the importance of economic policy dimensions as discussed in chapter 2.

52. An intermediate solution decided in 2008 was to set up *supervisory colleges** to discuss the risk profile of large cross-border financial institutions.

b) Incentives to risk-taking

On top of the regulatory loopholes, the pre-crisis period was characterized by regulations that did not prevent financial institutions from taking excessive risk and behaving in a procyclical way (or even in some instances that gave them incentives to behave in this way). There are many ways to correct the incentives of bankers, most of which have to do with the internal organization of banks: They involve strong risk-control departments, clear understanding by the bank's top management of the risks being taken, due diligence on clients' financial literacy to avoid the mis-selling of risky products, etc. It should have been the responsibility of shareholders to align the managers' incentives with their own interests and ensure that the right governance arrangements were in place. However, the crisis has revealed that shareholders have often been too short-sighted to care, meaning that supervisors have had to step in.

In order to correct incentives to leveraged risk-taking, several complementary routes have been followed.

- *Capital requirements*: A straightforward way to reduce the risk of bank failures is to raise capital requirements, to modulate them depending on liquidity mismatch between assets and liabilities, and to supplement them with limits on total leverage. The problem with tighter capital requirements, however, is their cost, since they amount to immobilizing capital that could usefully be employed elsewhere in the economy. A way not to waste capital would be Kashyap et al.'s (2008) *capital insurance** proposal. Under this scheme, banks would pay an ongoing premium to a "capital insurer" which would commit to inject capital into it in the event of a crisis. Candidate capital insurers would be long-term investors with a strong capital basis and no regulatory capital requirements, such as sovereign wealth funds, pension funds, or even governments (with the risk however that this new market would in fact be dominated by short-sighted investors, such as hedge funds). The advantage of this scheme would be to free up capital for productive use, rather than freezing it in banks' accounts. An objection is that too-big-to-fail, systemic financial institutions are already implicitly guaranteed by governments without paying an insurance premium. As noted by Alan Blinder, the analogy with insurance may not hold because the risk of a financial crisis is not diversifiable.[53] In addition, the issue of the trigger that would be used for the call of contingent capital is a delicate one: A rule-based regime might be too mechanical, but leaving room for discretion would inevitably give rise to intense lobbying.
- *Liquidity ratios*: The holding of liquid assets banks can easily draw on in times of stress is another way to prevent threats to the viability of

53. "The insurance premium is going to be extremely high, because you're making people pay in times when they don't want to pay," quoted in "Capital Ideas," *The Economist*, 28 August, 2008.

financial institutions in times of crisis. At the Pittsburgh summit it was agreed to add *liquidity ratios** to the existing or reformed capital ratios. The problem, however, is that a liquid market can suddenly dry up, prompting a wholesale increase in liquidity buffers. Hence liquidity requirements can be procyclical—see Lanoo and Casey (2005) for a discussion.

- *Counter-cyclical buffers*: Capital requirements force banks to raise capital or to extend less credit in troubled times, as already shown in box 8.2. To correct this procyclical feature of capital requirements, the introduction of time-variant *capital buffers** has been suggested: Under such a scheme, banks would be required to increase their regulatory capital or to set aside provisions[54] when credit accelerates and to dispose of them when losses have materialized, or are about to materialize. A scheme of this sort was put in place by the Bank of Spain in the 2000s after the country joined the euro, so as to mitigate the impact of the low euro area interest rates on domestic credit expansion.[55]
- *Compensation standards*: Performance-based bonuses have been identified as a source of risk-taking and procyclical behavior, since they have incited bank managements to inflate balance sheets in bull markets and sometimes to shrink them in bear markets. Accordingly, G20 leaders agreed to smooth them over longer time-spans and to introduce claw-back clauses (i.e., to cancel part of the bonus in the case of *ex post* underperformance), so that management bears responsibility for the full gamut of risks.
- *Taxation*: An alternative route to the strengthening of capital ratio is taxation. The concept of Pigouvian taxation discussed in chapter 7 can be applied to financial stability. This consists in taxing behavior that undermines overall financial stability. The idea here is not so much to make sure banks hold enough capital to avoid the risk of failure, but rather to make them internalize the externalities involved in financial risk-taking (Weder di Mauro, 2010; International Monetary Fund, 2010b). However, it is not clear that taxes on banks would provide better incentives than regulatory capital surcharges.

The correction of incentives is both the most natural route financial regulators can take and the most difficult one. It is the most natural because it can build on existing agreements, especially those put in place by the Basel Committee on Banking Supervision. But it is also the most difficult because appropriate financial regulation requires identifying externalities through which individual behavior threatens the stability of the financial system *as a whole*. In other words the goal is not only to make individual institutions more risk-conscious, but also to ensure that they price the risk they represent

54. This is called *through-the-cycle provisioning** or *dynamic provisioning**.

55. The difference between capital and provisions is that provisions dent operating profits, and are therefore more painful for shareholders.

to other institutions adequately. In addition, new capital requirements and liquidity rules add-up to generate significant capital needs, to be met on already stretched capital markets. They could have a non-negligible impact on the cost of capital. This is a major intellectual and operational challenge.

c) Market infrastructures

The same argument which holds for *actors* also holds for *markets*. Contrary to organized markets such as stock exchanges, which were tightly regulated to protect against market abuse, insider trading, etc., and were required to disclose information on prices and orders, *over-the-counter markets** (i.e., decentralized markets without a central counterpart or a clearing house) were not. At the apex of the crisis, no one could monitor the market for corporate *credit-default swaps** (CDSs),[56] where *counter-party risk** (i.e., the risk that the counter-party is unable to honor their contracts) could not be evaluated by market actors, which resulted in a drying up of the market. Consequently, it was decided at the G20 Pittsburgh summit that CDS markets should have a central counterparty to net out positions and that information on trades should be more readily accessible by supervisors.

d) The size and nature of banks

The failure of Lehman Brothers was made more dramatic due to the difficulty in identifying and compensating the bank's counterparties, since Lehman Brothers was not only "too big to fail," but also "too interconnected to fail." This points to the lack of comprehensive schemes to tackle large bankruptcies in the banking sector. Ironically, however, the crisis has spurred mergers and acquisitions that have led to an even larger number of systemically important institutions whose activities are spread over numerous countries.

One way to deal with this issue, proposed by the US Treasury (2009), is to admit that systemic banks will always be bailed out by governments, and, as the price for this insurance scheme, impose stricter and more conservative prudential standards on *Systemically Important Financial Institutions* (*SIFIs*)* in terms of capital and liquidity ratios and risk-management standards. Another possibility is to force the biggest banks to pre-plan their own demise by writing "living wills." This should not only make bank dissolutions easier and faster, but in the process of planning their own dissolutions banks would be encouraged to better track their exposure and possibly simplify their legal structure.[57]

56. Credit-default swaps (CDSs) are financial products that provide insurance against the risk of default of a private or public borrower. They are issued and traded by market participants. Lenders can use them to hedge against the risk of default of the borrower, and they can also be used for speculative purposes. See chapter 4.

57. Another proposal would be that, when a financial institution becomes insolvent, the regulator has the right to convert its debt into equity, see Snower (2009).

Some have gone further, suggesting that regulators should aim at making banks smaller, so that they can fail without creating trouble to the financial system. The issue here is whether banking involves significant economies of scale that justify banks growing in size, an issue on which the literature is inconclusive (Laeven and Levine, 2006). But the debate is as much about what banks do as about their size. In the spirit of the *Glass–Steagall Act** that regulated US banking between 1933 and 1999 and included a separation between commercial and investment banking, President Obama in January 2010 endorsed a proposal by Paul Volcker, the former chairman of the Federal Reserve. What he dubbed the "Volcker rule" was intended to prevent deposit-taking banks from owning, investing, or sponsoring hedge funds and to limit their ability to engage into proprietary trading (Obama, 2010).

At the extreme, proponents of the "narrow banking" claim that deposits should be protected from *any* risk, so that deposit-taking banks should only hold safe securities on their balance sheet. However, the business boundaries of the 1930s cannot simply be replicated, and this ignores the fact that the transformation of short, liquid deposits into longer-term, riskier investments fulfils a macroeconomic role as illustrated by the Diamond–Dybvig model discussed in chapter 4.[58]

e) Organize macro-prudential supervision and regulation

The concept of *macro-prudential regulation and supervision** dates back to the 1970s and has long been championed by the Bank for International Settlements,[59] but it gained traction only after the crisis erupted. In a nutshell, it consists of supplementing monetary policy by another instrument that allows the authority in charge to recommend or enforce measures that prevent financial instability. The discussion in the 2000s of asset prices and monetary policy (see chapter 4) was, inadvertently, about macro-prudential regulation. The lack of consensus on how to implement macro-prudential regulation[60] suggests avoiding a pure rule-based approach and giving a degree of discretion to a supervisory authority. This can also be viewed as a learning process.

This leads us to the *who* question. There is consensus that macro-prudential oversight should involve central banks, because they are technically equipped both for macro-financial and for micro-financial analysis, and because they should anyway be prepared to act as lenders of last resort when systemic risks materialize (see chapter 4). But macro-prudential regulation also requires a bird's-eye view of systemic risk in the global financial system—something the EU achieved in 2009 by creating a *European Systemic Risk Board** led by the ECB as proposed by the De Larosière report, and a role which has been

58. Wallace (1996) provides a formal treatment of this question.

59. See C. Borio "The macroprudential approach to regulation and supervision," *Vox*, 14 April 2009, and Borio (2003).

60. See the BIS 79th Annual Report, Goodhart (2009), and Repullo et al. (2009).

taken over by the Federal Reserve in the US. However, the feedback from macro-variables to micro-regulations and standards raises tricky questions. How can a central bank, or a similar body, enforce regulatory changes in spite of not having competence on regulatory matters? The vagaries of the Stability and Growth Pact, another macro-based regulatory framework (chapter 3 and supra), should caution against excessive faith in output gaps and sophisticated through-the-cycle incentive schemes.

Finally comes the *what for* question. Giving central banks a macro-prudential instrument implies that they know when to use it instead of using their interest-rate instrument. A simple answer is to say that the interest rate should be used to target consumer-price inflation and the macro-prudential instrument to target asset-price inflation or credit growth. However, this may lead to situations when a central bank does one thing with the right hand and another with the left. If central banks are to be given a second objective and a second instrument, this calls for an in-depth reexamination of the pre-existing policy consensus and the elaboration of a new policy doctrine.

While the tasks of central banks have been vastly expanded during the crisis, their constitutional mandate and governance model have not been revisited. Even though some of these tasks are discontinued when the crisis is over, this raises a dilemma. If the central bank retains a purely advisory role, it risks losing its credibility by being held responsible for outcomes which in effect it cannot control. Think of housing bubbles: A central bank responsible for financial stability (say, the ECB in its role as chair of the European Systemic Risk Board) may urge governments to take regulatory or tax action to cool down the housing market but damage its own credibility if they do not comply. However, if the instruments to enforce financial stability are devolved to it, the central bank will have many instruments in its hands and many objectives to achieve. In the absence of a clear mission statement, it will soon experience conflicts of interest and make mistakes. And it may well be challenged politically as being too powerful and in control of instruments whose use requires parliamentary oversight. This would be all the more likely because central banks could be, through the setting of cyclical capital buffers, at the origin of the need for bank bail-outs. During the crisis itself the independence of the Fed started being questioned by Congress. Giving central banks too many objectives could eventually result in a loss of independence and hence less capacity to achieve price stability.[61]

8.3.2 The macroeconomic policy regime

In the wake of the crisis the macroeconomic focus was on remedial action rather than longer-term reform. As regards national measures, global summits and other international gatherings put emphasis first on stimulus

61. On macro-prudential supervision and central bank independence, see John Taylor, "Fed needs better performance, not powers," *Financial Times*, 10 August 2009.

measures and financial regulation, then on exit strategies and on resources and governance of international organizations. However, a number of macro-issues emerged.

a) A new view on rules versus discretion?

As developed in section 8.2, rule-based policymaking was largely put aside during the crisis. The EU fiscal policy is a case in point. The European Stability and Growth Pact (SGP) was constructed on the assumption that the general government deficit in a given country would move up and down within a limited range along the business cycle, but in 2009–10 the EU fiscal deficit widened by 5% of EU GDP and more than 20 of the 27 EU member states were considered by the European Commission to have an excessive deficit. The SGP includes an escape clause for exceptional and temporary deficits above the 3% of GDP threshold, but it does not set out the principles to be applied on such occasions. As a matter of fact, for most EU countries, the European commission considered that, although circumstances were "exceptional," large budgetary deficits were not "temporary," so the escape clause could not apply.

On the one hand, contingent policy rules in case of crisis are difficult to specify because all crises are different, and because unexpected shocks and rapidly unfolding events are best addressed by discretionary action. On the other hand, letting policymakers depart too easily from rules they have themselves defined undermines the credibility and the very effectiveness of these rules—remember the value of commitment, as explained in chapter 2. Hence the need for policy rules to include well-formulated *escape clauses** in order to make room for temporary discretion and centralization, but also to be specific on which circumstances qualify as extraordinary.

Escape clauses are no "free lunch": As illustrated by the experience with fixed exchange-rate regimes, to leave open the possibility of departure from the stated rules leads markets to price in the corresponding risk—for example, through higher risk premiums on government debt. But it may be a cost worth paying. Similarly, some countries, most notably Germany, have concluded that tighter fiscal rules in normal times are a desirable quid pro quo for flexibility in crisis times and have reformed their constitution accordingly.

b) International coordination and surveillance

International coordination may not be confined to prudential issues, since macroeconomic factors have also played a role in the crisis. However, it took time before G20 statements started to address global imbalances, and even more time before they dared to address monetary policies and the international monetary system.

The reason is that governments (i) do not agree on where the responsibility for the crisis lies and (ii) are reluctant to commit to abiding by rules that

would put constraints on their economic policy decisions. As illustrated by the discussion on the "global savings glut" of section 8.1, global imbalances can be considered the result of either excessively low saving in the US or excessively high net saving abroad; or they may result from emerging countries' willingness to self-insure against future sudden stops in capital inflows by accumulating foreign-exchange reserves; or they can result from monetary authorities in some large economies refraining from letting their currency appreciate through foreign-exchange interventions, when they experience current-account surpluses and/or large capital inflows. In fact, global imbalances are a general equilibrium outcome whose policy roots are hard to pin down (see the model of Blanchard et al.,2005, presented in chapter 5, and Obstfeld and Rogoff, 2009, for a discussion).

At the Pittsburgh summit of September 2009, G20 leaders established a "Framework for Strong, Sustainable, and Balanced Growth" and asked finance ministers and governors to "set out objectives, put forward policies to achieve these objectives, and together assess [their] progress." This revival of coordination contrasts with at least 20 years of emphasis on independent national policymaking. In the 1990s and the 2000s, exchange-rate surveillance, a core mission of the IMF, could not be exercised effectively (Independent Evaluation Office of the IMF, 2006). The IMF was neither able to influence US policy nor even to express a public view on China's exchange-rate policy.

For IMF surveillance to be credible it requires even-handedness, which in turn calls for a reform of the Fund's governance. Currently the US retains veto power on all important decisions (which require an 85% majority) and Europe is globally overrepresented with about a third of total voting rights. China, which before the crisis had fewer votes than France, and India, which ranked behind Italy, could not accept a stronger IMF unless this came with a major power shift. Even when this is achieved, whether or not surveillance can constrain national policies will remain an open issue. The European experience discussed in chapters 3–5 is not very encouraging in this respect.

c) Self-insurance or collective insurance?

One reason why East Asian countries went into current account surpluses in the 2000s was their desire to accumulate foreign-exchange reserves in order to be able to cushion capital-flow reversals. Their experience during the crisis of 1997–98 and what they perceived as a western bias in IMF decisions led them to insure themselves through reserve accumulation instead of relying on IMF support in the event of a balance-of-payments crisis. Such self-insurance behavior was costly in at least two respects: The fiscal cost of sterilizing the induced rise in domestic liquidity, and the political cost of being accused of currency manipulation by trading partners which let their exchange rates float freely.

Together with IMF governance reform, a series of G20 decisions—the tripling of IMF resources from 250 to 750 billion dollars, the weaker policy

conditionality of its programs, and the introduction of a new, unconditional *Flexible Credit Line** (FCL)—were intended to address these concerns and relax the constraint of reserve accumulation.[62] However, as noted by Maurice Obstfeld (2009), the world still lacks a global lender of last resort. The IMF would be a natural candidate, but in 2009 it was the swap lines provided bilaterally (and in a fully discretionary way) by the US Fed that provided to emerging countries like Korea the comfort they needed. The IMF's FCL was extended to Poland, Mexico, and Columbia only, and none of these three countries drew on it during the crisis. Post-crisis history will tell whether a universal scheme for liquidity provision will emerge.

8.3.3 Conclusion

As discussed in section 8.1, there are still several explanations to the crisis. Even though they are not mutually exclusive, they result in different sets of policy recommendations, all of which combine the overhaul of financial regulation, supervisory reform, changes in the monetary policy framework, and some of which also involve reform of the international monetary system and the remit and governance of international organizations. The G20, relying on specialized international bodies such as the IMF and FSB, as well as on national and regional authorities, has addressed some of them. Some crucial issues, however, have been left unaddressed, both on the regulatory and the macro-financial fronts.

a) The remaining regulatory challenge

Three issues stand out as unresolved challenges:

- First, *moral hazard* was magnified by the crisis because of the post-Lehman G7 decision not to let any further financial institution of systemic significance collapse. Large and interconnected institutions, as well as their shareholders, now know that they will be rescued if threatened with default. This entails a significant danger of excessive risk-taking—the very same danger the whole apparatus of regulation is intended to avert. However, there is no limit to bank size, the failing banks' creditors have not been penalized in the rescue operations, and even shareholders have not borne the full brunt of their responsibility. This is in striking contrast with the Asian crisis of the late 1990s, when "private sector involvement" (meaning sharing the losses) was the name of the game.
- Second, the *desirable size* of the financial sector has barely been discussed. As illustrated by its employees' generous compensation

62. The FCL was extended in 2009 to Mexico, Poland, and Colombia for a total amount of 77.9 billion dollars.

(Philippon and Reshef, 2009), the banking sector seems to have succeeded in capturing a rent, which implies that allocating more and more resources to this sector could end up being detrimental to overall economic efficiency. Furthermore, as illustrated by the Icelandic meltdown, a large financial sector in a small- or medium-sized country entails the risk of incurring proportionally very large public-finance costs in the event of a bail-out. Lord Turner, the head of the UK's Financial Services Authority, has advocated taxing financial institutions in order to tame their development, and the idea has been taken up by the IMF (2010b) in a proposal to the G20. However, this proposal competes with other motives to tax the financial system, such as the compensation of costs incurred during the crisis, the creation of insurance funds, or Pigouvian taxation, as discussed above, and proposals for financial transactions taxes intended to provide resources for the financing of global public goods (see chapter 7).

- Last but not least, the *trade-off* between financial stability and the cost of capital has not really been addressed. Many of the financial stability measures on the official agenda will result in increasing the cost of capital. This is for example the case with capital adequacy ratios: Increasing them will make financing more costly for nonfinancial companies, with adverse consequences for capital expenditure and technological innovation. The question here is what price society is willing to pay as a counterpart to financial stability: Is a more unstable economy acceptable, if it is the condition for faster growth? This fundamental question was at the core of the cost–benefit assessment of the new "Basel 3" framework for bank supervision. The answer, which relates to collective preferences, is unlikely to be the same across countries. This suggests that regulatory discrepancies are here to stay.

b) Open macro-financial issues

Turning to the macro-financial dimensions, three items deserve mention:

- *Stress-testing economic policy*. The crisis has been a reminder that economic policy involves a strong risk-management dimension. This was understood before the crisis by corporations (although risk-management measures were admittedly too crude), but hardly at all by governments. Governments do not assess risk properly (remember the example of public health tests and the analysis of fiscal-policy behavior of chapter 2) and they seldom disclose margins of error for their own evaluations. More importantly, they do not implement the kind of *stress-testing** that is required of financial institutions , i.e., assessments of the robustness of their solvency to extremely unlikely combinations of events: What if the stock market crashes by more than x%, oil prices rise by y%, and recovery rates on loans are less than z%?

Admittedly, stress-testing is even more difficult for a government than it is for a company, because it cannot be conceived as a partial equilibrium exercise and requires an assessment of the robustness of the whole economic and financial complex. However, economic policymakers should learn from robustness assessments such as the one routinely undertaken in complex industrial and IT systems.

- *Which new framework for macroeconomic policy*? As described in chapter 4, by the mid-2000s many countries (though not all) were converging on a monetary policy framework that gave a primary role to flexible inflation targeting and on a dismissive view of the stabilization role of fiscal policy. Even the central banks whose mandate encompassed growth (like the Fed), or those that claimed to preserve a role for the monetary aggregates (like the ECB) were de facto moving in the direction of inflation targeting, and even in the traditionally fiscally activist countries, tax and spending changes scarcely responded to cyclical developments. Furthermore, the macroeconomic policy framework ignored financial stability dimensions. The crisis has called into question this framework, but no consensus has yet emerged on its reform or replacement. In a much-commented-upon paper, Blanchard et al. (2010) have proposed strengthening the ability of macroeconomic policy to respond to crises through raising the inflation target (to create more monetary policy space, see chapter 4), introducing stronger, nonlinear automatic stabilizers (to make fiscal policy more responsive, see chapter 3), and making room for macro-prudential policies; but discussion thus far has been dominated by the central bankers' outcry about the suggestion of having a somewhat higher inflation target.
- *What reform of the international monetary system*? For those who believe that the Chinese current-account surplus and the US deficit played an important role in creating the conditions for financial instability, the key question is how to engineer a reduction of these imbalances. Beyond the reform of surveillance mentioned in the previous section, discussion has started on the reform of international monetary arrangements. In March 2009, Chinese central bank governor Zhou Xiaochuan (2009) called for the creation of "an international reserve currency that is disconnected from individual nations and is able to remain stable in the long run, thus removing the inherent deficiencies caused by using credit-based national currencies."

This open challenge to the monopolistic role of the US dollar as an international currency—and the resulting lack of incentives for US discipline—was an invitation to re-open an international monetary discussion that had been stagnant since the demise of the Plaza–Louvre agreements (see chapter 5). However, the obstacles to redefinition of the global rules

of the monetary game are even more formidable than those to a strengthening of surveillance. The SDR could play a greater role as a reserve of value but could hardly replace the dollar as the dominant international currency. The Chinese renminbi will surely play a role at some point in time, but full convertibility is a precondition. At any rate, there is currently no challenger to the international role of the dollar (Pisani-Ferry and Posen, 2009; Eichengreen, 2009).

References

Acharya, V., and M. Richardson (eds) (2009), *Restoring Financial Stability: How to Repair a Failed System*, Wiley.

Adrian, T., and H.-S. Shin (2008), "Liquidity and Leverage," *Federal Reserve Bank of New York Staff Reports* no. 328, May.

Adrian, T., and H.-S. Shin (2009), "The Shadow Banking System: Implications for Financial Regulation," *Federal Reserve Bank of New York Staff Reports* no. 382, July.

Bank for International Settlements (2008), *78th Annual Report.*

Bebchuk, L. (2009), *Written Testimony Submitted to the Committee on Financial Services,* US House of Representatives, 11 June.

Bénassy-Quéré, A., Y. Decreux, L. Fontagné, and D. Khoudour-Castéras (2009), "Economic crisis and global supply chains," *CEPII working paper* no. 2009–15, July.

Bernanke, B. (2002), "Deflation: Making Sure 'IT' Doesn't Happen Here," Remarks Before the National Economists Club, 21 November.

Bernanke, B. (2005), "The Global Savings Glut and the US Current Account Deficit," Sandridge Lecture, Virginia Association of Economics, 10 March.

Bernanke, B. (2009), "The Crisis and the Policy Responses," At the Stamp Lecture, London School of Economics, London, 13 January.

Bernanke, B. (2010), "Monetary Policy and the Housing Bubble," Speech at the American Economic Association Meeting, 3 January.

Berndt, A., and A. Gupta (2009), "Moral Hazard and Adverse Selection in the Originate-to-Distribute Model of Bank Credit," *Journal of Monetary Economics*, 56, pp. 725–43.

Blanchard, O., and G.- M. Milesi-Ferretti (2009), "Global Imbalances: In Midstream?," *IMF Staff Position Note* 09/29, December.

Blanchard, O., and J. Simon (2001), "The Long and Large Decline in US Output Volatility," *Brookings Papers on Economic Activity* 2001–1, pp. 135–64.

Blanchard, O., and J. Wolfers (2000), "The Role of Shocks and Institutions in the Rise of European Unemployment: The Aggregate Evidence," *Economic Journal*, 110, pp. 1–33.

Blanchard, O., F. Giavazzi, and F. Sa (2005), "International Investors, the US Current Account, and the Dollar," *Brookings Papers on Economic Activity*, 2005–1, pp. 1–49.

Blanchard, O., C. Cottarelli, A. Spilimbergo, and S. Symansky (2008), "Fiscal Policy for the Crisis," *IMF Staff Position Note* 08/01, December.

Blanchard, O., G. Dell' Ariccia, and P. Mauro (2010), "Rethinking Macroeconomic Policy," *IMF Staff Position Note* 10/03, February.

Blankfein, L. (2009), "Do Not Destroy the Essential Catalyst of Risk," *The Financial Times*, 8 February.
Boeri, T., L. Bovenberg, B. Coeuré, and A. Roberts (2006), "Dealing with the New Giants: Rethinking the Role of Pension Funds," *CEPR/ICMB Geneva Report on the World Economy* no 8.
Borio, C. (2003), "Towards a Macroprudential Framework for Financial Supervision and Regulation?" *CESifo Economic Studies*, 49, pp 181–216.
Brender, A., and F. Pisani (2009), *Globalised Finance and Its Collapse*, Dexia, available at https://www.dexia-am.com/globalisedfinance/Globalisedfinance.pdf.
Broda, C., P. Ghezzi, and E. Levy-Yeyati (2009), "The New Global Balance: Financial De-globalisation, Savings Drain, and the US Dollar," *Vox*, 22 May.
Brown, G. (2005), *Speech at the CBI Annual Conference*, 28 November.
Buiter, W. (2007), "Lessons from the 2007 Financial Crisis," *CEPR Policy Insight 18*, Centre for Economic Policy Research, December.
Caballero, R. (2009), "The 'Other' Imbalance and the Financial Crisis", Paolo Baffi Lecture, mimeo, December.
Caballero, R., and A. Krishnamurthy (2009), "Global Imbalances and Financial Fragility," *American Economic Review*, 99, 584–88.
Caballero, R., E. Farhi, and P.-O. Gourinchas (2008), "An Equilibrium Model of Global Imbalances and Low Interest Rates," *American Economic Review*, 98, pp. 358–93.
Cerra, V., and S. C. Saxena (2008), "Growth Dynamics: The Myth of Economic Recovery," *American Economic Review*, 98, pp. 439–57.
Coval, J., J. Jurek, and E. Stafford (2009), "The Economics of Structured Finance," *Journal of Economic Perspectives*, 23, pp. 3–25.
De Larosière, J. (2009), *Report of the High-Level Group on Financial Supervision in the EU*, European Commission, February.
De Mandeville, B. (1714 [1989]), *The Fable of the Bees: Or Private Vices, Publick Benefits*, Penguin Classics.
Dooley, M., D. Folkerts-Landau, and P. Garber (2003), "An Essay on the Revived Bretton Woods System," *NBER Working Paper* no. 9971.
Eichengreen, B. (2009), "The Dollar Dilemma," *Foreign Affairs*, 88, pp. 53–68.
Eichengreen, B., and K. O'Rourke (2009), *A Tale of Two Depressions*, available at www.voxeu.org, June, updated in *What Do the New Data Tell Us?*, March 2010, available on www.voxeu.org.
European Commission (2009), *Public Finance in EMU 2009*, June.
Engel, C., and J. Rogers (2006), "The US Current Account Deficit and the Expected Share of World Output," *Journal of Monetary Economics*, 53, pp. 1063–93.
Feldstein, M. (2009), "The Case for Fiscal Stimulus," Project Syndicate, January, available at www.project-syndicate.org.
Financial Services Authority (2009), *The Turner Review: A Regulatory Response to the Global Banking Crisis*, Financial Services Authority, London, March.
Friedman, M., and A. Schwartz (1963), *A Monetary History of the United States*, 1867–1960, Princeton University Press (for the National Bureau of Economic Research).
Goodhart, C. (2009), "Procyclicality and Financial Regulation," *Revista de Estabilidad Financiera* no. 16, Banco de España, May.
Gorton, G. (2008), "The Panic of 2007," in *Maintaining Stability in a Changing Financial System*, Proceedings of the 2008 Jackson Hole Conference, Federal Reserve Bank of Kansas City.

Gorton, G. (2009a), "The Subprime Panic," *European Financial Management*, 15, pp. 10–46.
Gorton, G. (2009b), "Slapped in the Face by the Invisible Hand: Banking and the Panic of 2007," paper prepared for the Federal Reserve Bank of Atlanta's 2009 Financial Markets Conference: Financial Innovation and Crisis, 11–13 May.
Gourinchas, P.-O., and H. Rey (2007), "From World Banker to World Venture Capitalist: US External Adjustment and the Exorbitant Privilege," in Clarida, R. (ed.), *G7 Current Account Imbalances: Sustainability and Adjustment*, National Bureau of Economic Research, pp. 11–66.
Greenspan, A. (2004), "The Mortgage Market and Consumer Debt," Remarks at America's Community Bankers Annual Convention, Washington, D.C., 19 October.
Greenspan, A. (2005), Testimony before the Committee on Banking, Housing, and Urban Affairs, US Senate, 16 February.
Haldane, A. (2009), "Rethinking the Financial Network," Speech at the Financial Student Association, Amsterdam, 28 April.
Hellwig, M. (2008), "The Causes of the Financial Crisis," *CESifo Forum 4/2008*, pp. 12–21.
Hildebrand, P. (2008), "Is Basel II Enough? The Benefits of a Leverage Ratio," Financial Markets Group Lecture, London School of Economics, 15 December.
Holmström, B. (2008), "Discussion of 'The Panic of 2007' by Gary Gorton," in *Maintaining Stability in a Changing Financial System*, Proceedings of the 2008 Jackson Hole Conference, Federal Reserve Bank of Kansas City.
Horton, M., M. Kumar, and P. Mauro (2009), "The State of Public Finances: A Cross-Country Fiscal Monitor," *IMF Staff Position Note* 09/21, July.
Hoshi, T., and A. Kashyap (2008), "Will the US Bank Recapitalization Succeed? Lessons from Japan," *Working Paper* no. 14401, National Bureau of Economic Research.
Independent Evaluation Office of the IMF (2006), *Evaluation of IMF Multilateral Surveillance*, IEO Report, September.
International Monetary Fund (2008), *Global Financial Stability Report*, October.
International Monetary Fund (2009), "Fiscal Implications of the Global Economic and Financial Crisis," *Staff Position Paper* 09/13, June.
International Monetary Fund (2010a), *Global Financial Stability Report*, April.
International Monetary Fund (2010b), *A Fair and Substantial Contribution by the Financial Sector*, Interim Report to the G20, April.
Kashyap, A., R. Rajan, and J. Stein (2008), "Rethinking Capital Regulation," in *Maintaining Stability in a Changing Financial System*, Proceedings of the 2008 Jackson Hole Conference, Federal Reserve Bank of Kansas City.
King, M. (2009), Speech at the CBI Dinner, Nottingham, 20 January.
Kroszner, R. (2009), "Central Banks Must Time a 'Good Exit'," *The Financial Times*, 11 August.
Krugman, P. (1998a), "Will Asia Bounce Back?" March, available on http://web.mit/krugman/www.
Krugman, P. (1998b), "The Confidence Game: How Washington Worsened Asia's Crash," *The New Republic*, 5 October.
Laeven, L., and R. Levine (2006) "Is there a Diversification Discount in Financial Conglomerates?" *CEPR Discussion Paper* no. 5121, July.
Laeven, L., and F. Valencia (2008), "Systemic Banking Crises: A New Database," *IMF*

Working Paper 08/224, November.
Lannoo, K., and J.-P. Casey (2005), "Capital Adequacy vs. Liquidity Requirements in Banking Supervision in the EU," *CEPS Policy Brief* no. 84, October.
Megginson, W., and J. Netter (2001), "From State to Market: A Survey of Empirical Studies on Privatization," *Journal of Economic Literature*, 39, pp. 321–89.
Meier, A. (2009), "Panacea, Curse or Nonevent? Unconventional Monetary Policy in the United Kingdom," *IMF Working Paper* 09/163, August.
Mendoza, E., V. Quadrini, and J.-V. Ríos Rull (2009), "Financial Integration, Financial Development and Global Imbalances," *Journal of Political Economy*, 117, pp. 371–416.
Obama, B. (2009), Address at George Mason University in Fairfax, Virginia, 8 January.
Obama, B. (2010), Remarks on Financial Reform, January.
Obstfeld, M. (2009), "Lenders of Last Resort in A Globalized World," University of California, Berkeley, mimeo, June.
Obstfeld, M. and K. Rogoff (2005), "The Unsustainable US Current Account Position Revisited," in Clarida, R. (ed.), *G7 Current Account Imbalances: Sustainability and Adjustment*, National Bureau of Economic Research, pp. 339–76.
OECD (2010), "The Impact of the Economic Crisis on Potential Output", Working party no. 1 on Macroeconomic and Structural Policy Analysis, ECO/CPE/WP1(2010)3, February, Paris: Organization for Economic Cooperation and Development.
Orphanides, A. (2004), "Monetary Policy in Deflation: The Liquidity Trap in History and Practice," *The North American Journal of Economics and Finance*, 15, pp. 101–24.
Panetta, F., and P. Angelini (eds) (2009), "Financial Pro-cyclicality: Lessons from the Crisis," *Questioni di Economia e Finanza*, 44, Banca d'Italia, April.
Panetta, F., T. Faeh, G. Grande, C. Ho, M. King, A. Levy, F. Signoretti, M. Taboga, and A. Zaghini (2009), "An Assessment of Financial Sector Rescue Programmes," *BIS Papers* no. 48, Bank for International Settlements.
Persaud, A. (2009), "Macro-Prudential Regulation: Fixing Fundamental Market (and Regulatory) Failures," paper prepared for the Bruegel-Cepii-Icrier conference on International Cooperation in Times of Global Crisis: Views from G20 Countries, New Delhi, 14–15 September.
Philippon, T., and A. Reshef (2009), "Wages and Human Capital in the US Financial Industry: 1909–2006," *NBER Working Paper* 14644, January.
Pisani-Ferry, J., and A. Posen (2009), *The Euro at Ten: The Next Global Currency?*, Bruegel/Peterson Institute for International Economics.
Pisani-Ferry, J., and A. Sapir (2010), "Banking Crisis Management in the EU: an Early Assessment," *Economic Policy*, 62, pp. 343–73.
Pisani-Ferry, J., and B. Van Pottelsberghe (2009), "Handle with Care: Post-Crisis Growth in the EU," *Bruegel Policy Brief* no. 2009/02, April.
Reinhart, C., and K. Rogoff (2009a), "The Aftermaths of Financial Crises," *American Economic Review*, 99, pp. 466–72.
Reinhart, C., and K. Rogoff (2009b), *This Time Is Different: Eight Centuries of Financial Folly*," Princeton University Press.
Repullo, R., and J. Suarez (2008), "The Procyclical Effects of Basel II," *CEPR Discussion Paper* no. 6862, June.
Repullo, R., J. Saurina, and C. Trucharte (2009), "Mitigating the Procyclicality of Basel II," *CEMFI Working Pape*r no. 0903, July.

Rodrik, D. (2006), "The Social Cost of Foreign Exchange Reserves," *International Economic Journal*, 20, pp. 253–66.
Rodrik, D. (2009), "A Plan B for Global Finance," *The Economist*, 12 March.
Romer, C. (1999), "Change in Business Cycles: Evidence and Explanations," *Journal of Economic Perspectives*, 13, pp. 23–44.
Smith, A. (1776 [1977]), *An Inquiry into the Nature and Causes of the Wealth of Nations*, University of Chicago Press.
Snower, D. (2009), "The Impact of the Global Financial Crisis on Europe and Europe's Responses," presentation at the Asia-Europe Economic Forum conference, Kiel, 7–8 July.
Taleb, N. N. (2007), *The Black Swan: The Impact of the Highly Improbable*, Random House.
Taylor, J. (2008), *Getting Off Track: How Government Actions and Interventions Caused, Prolonged, and Worsened the Financial Crisis*, Hoover Institution Press, Stanford, March.
US Treasury (2009), *Financial Regulatory Reform: A New Foundation*, US Government Report, June.
Volcker, P. (2009), *Statement before the Committee on Banking and Financial Services*, US House of Representatives, 24 September.
Von Hagen, J. (2009), "The Monetary Mechanics of the Crisis," *Bruegel Policy Contribution* no. 2009/08, August.
Wallace, N. (1996), "Narrow Banking Meets the Diamond–Dybvig Model," *Federal Reserve Bank of Minneapolis Quarterly Review*, 20, pp. 3–13.
Warnock, F., and V. Cacdac Warnock (2009), "International Capital Flows and US Interest Rates", *Journal of International Money and Finance*, 28, pp 903–19.
Weder di Mauro, B. (2010), "Taxing Systemic Risk: Proposal for a Systemic Risk Levy and a Systemic Risk Fund," paper presented at Bundesbank seminar, January.