Cursuswijzer Veranderende Verhoudingen in de Wereldeconomie

HOVO 2017

Collegeschema (Maandag 15.15-17.00)

<table>
<thead>
<tr>
<th>18/10</th>
<th>Alle welvaart komt uit Utrecht.</th>
<th>Roelfsema</th>
</tr>
</thead>
<tbody>
<tr>
<td>25/10</td>
<td>Globalisering, technologische ontwikkeling en de Future of Work.</td>
<td>Roelfsema</td>
</tr>
<tr>
<td>2/10</td>
<td>Financiële crises en mondiaal financiële Integratie.</td>
<td>Roelfsema</td>
</tr>
<tr>
<td>9/10</td>
<td>Opkomende markten en economische ontwikkeling</td>
<td>de Bell</td>
</tr>
<tr>
<td>23/10</td>
<td>Mondiale inkomensongelijkheid en migratie.</td>
<td>de Bell</td>
</tr>
<tr>
<td>30/10</td>
<td>De Toekomst van de Europese Unie, Trump, en regionale integratie.</td>
<td>Roelfsema</td>
</tr>
</tbody>
</table>

Reader


**Week 5:** Milanovic, B. (2011). Global inequality: from class to location, from proletarians to migrants.

Prepared for the Handbook of Economic Growth edited by Philippe Aghion and Steve Durlauf. We thank the editors for their patience and Leopoldo Fergusson, Pablo Querubín and Barry Weingast for their helpful suggestions. The views expressed herein are those of the author(s) and not necessarily those of the National Bureau of Economic Research.

©2004 by Daron Acemoglu, Simon Johnson, and James Robinson. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.
Institutions as the Fundamental Cause of Long-Run Growth
Daron Acemoglu, Simon Johnson, and James Robinson
NBER Working Paper No. 10481
May 2004
JEL No. N11, N13, N15, N16, N17, O10, P10, P17

ABSTRACT

This paper develops the empirical and theoretical case that differences in economic institutions are the fundamental cause of differences in economic development. We first document the empirical importance of institutions by focusing on two "quasi-natural experiments" in history, the division of Korea into two parts with very different economic institutions and the colonization of much of the world by European powers starting in the fifteenth century. We then develop the basic outline of a framework for thinking about why economic institutions differ across countries. Economic institutions determine the incentives of and the constraints on economic actors, and shape economic outcomes. As such, they are social decisions, chosen for their consequences. Because different groups and individuals typically benefit from different economic institutions, there is generally a conflict over these social choices, ultimately resolved in favor of groups with greater political power. The distribution of political power in society is in turn determined by political institutions and the distribution of resources. Political institutions allocate de jure political power, while groups with greater economic might typically possess greater de facto political power. We therefore view the appropriate theoretical framework as a dynamic one with political institutions and the distribution of resources as the state variables. These variables themselves change over time because prevailing economic institutions affect the distribution of resources, and because groups with de facto political power today strive to change political institutions in order to increase their de jure political power in the future. Economic institutions encouraging economic growth emerge when political institutions allocate power to groups with interests in broad-based property rights enforcement, when they create effective constraints on power-holders, and when there are relatively few rents to be captured by power-holders. We illustrate the assumptions, the workings and the implications of this framework using a number of historical examples.

Daron Acemoglu
Department of Economics
MIT
50 Memorial Drive
Cambridge, MA 02142
and NBER
daron@mit.edu

James Robinson
Departments of Economics and Political Science
University of California, Berkeley
Berkeley, CA 94720

Simon Johnson
Sloan School of Management
MIT
50 Memorial Drive
Cambridge, MA 02142
and NBER
sjohnson@mit.edu
1 Introduction

1.1 The question

The most trite yet crucial question in the field of economic growth and development is: Why are some countries much poorer than others? Traditional neoclassical growth models, following Solow (1956), Cass (1965) and Koopmans (1965), explain differences in income per capita in terms of different paths of factor accumulation. In these models, cross-country differences in factor accumulation are due either to differences in saving rates (Solow), preferences (Cass-Koopmans), or other exogenous parameters, such as total factor productivity growth. More recent incarnations of growth theory, following Romer (1986) and Lucas (1988), endogenize steady-state growth and technical progress, but their explanation for income differences is similar to that of the older theories. For instance, in the model of Romer (1990), a country may be more prosperous than another if it allocates more resources to innovation, but what determines this is essentially preferences and properties of the technology for creating ‘ideas’.

Though this theoretical tradition is still vibrant in economics and has provided many insights about the mechanics of economic growth, it has for a long time seemed unable to provide a fundamental explanation for economic growth. As North and Thomas (1973, p. 2) put it: “the factors we have listed (innovation, economies of scale, education, capital accumulation etc.) are not causes of growth; they are growth” (italics in original). Factor accumulation and innovation are only proximate causes of growth. In North and Thomas’s view, the fundamental explanation of comparative growth is differences in institutions.

What are institutions exactly? North (1990, p. 3) offers the following definition: “Institutions are the rules of the game in a society or, more formally, are the humanly devised constraints that shape human interaction.” He goes on to emphasize the key implications of institutions since, “In consequence they structure incentives in human exchange, whether political, social, or economic.”

Of primary importance to economic outcomes are the economic institutions in society such as the structure of property rights and the presence and perfection of markets. Economic institutions are important because they influence the structure of economic

---

1 Although some recent contributions to growth theory emphasize the importance of economic policies, such as taxes, subsidies to research, barriers to technology adoption and human capital policy, they typically do not present an explanation for why there are differences in these policies across countries.
incentives in society. Without property rights, individuals will not have the incentive to invest in physical or human capital or adopt more efficient technologies. Economic institutions are also important because they help to allocate resources to their most efficient uses, they determine who gets profits, revenues and residual rights of control. When markets are missing or ignored (as they were in the Soviet Union, for example), gains from trade go unexploited and resources are misallocated. Societies with economic institutions that facilitate and encourage factor accumulation, innovation and the efficient allocation of resources will prosper.

Central to this chapter and to much of political economy research on institutions is that economic institutions, and institutions more broadly, are endogenous; they are, at least in part, determined by society, or a segment of it. Consequently, the question of why some societies are much poorer than others is closely related to the question of why some societies have much “worse economic institutions” than others.

Even though many scholars including John Locke, Adam Smith, John Stuart Mill, Douglass North and Robert Thomas have emphasized the importance of economic institutions, we are far from a useful framework for thinking about how economic institutions are determined and why they vary across countries. In other words, while we have good reason to believe that economic institutions matter for economic growth, we lack the crucial comparative static results which will allow us to explain why equilibrium economic institutions differ (and perhaps this is part of the reason why much of the economics literature has focused on the proximate causes of economic growth, largely neglecting fundamental institutional causes).

This chapter has three aims. First, we selectively review the evidence that differences in economic institutions are a fundamental cause of cross-country differences in prosperity. Second, we outline a framework for thinking about why economic institutions vary across countries. We emphasize the potential comparative static results of this framework and also illustrate the key mechanisms through a series of historical examples and case studies. Finally, we highlight a large number of areas where future theoretical and empirical work would be very fruitful.

1.2 The Argument

The basic argument of this chapter can be summarized as follows:

1. Economic institutions matter for economic growth because they shape the incentives of key economic actors in society, in particular, they influence investments in physical and human capital and technology, and the organization of production. Al-
though cultural and geographical factors may also matter for economic performance, differences in economic institutions are the major source of cross-country differences in economic growth and prosperity. Economic institutions not only determine the aggregate economic growth potential of the economy, but also an array of economic outcomes, including the distribution of resources in the future (i.e., the distribution of wealth, of physical capital or human capital). In other words, they influence not only the size of the aggregate pie, but how this pie is divided among different groups and individuals in society. We summarize these ideas schematically as (where the subscript $t$ refers to current period and $t+1$ to the future):

$$
\text{economic institutions}_t \implies \left\{ \begin{array}{l}
\text{economic performance}_t \\
\text{distribution of resources}_{t+1}
\end{array} \right.
$$

2. Economic institutions are endogenous. They are determined as collective choices of the society, in large part for their economic consequences. However, there is no guarantee that all individuals and groups will prefer the same set of economic institutions because, as noted above, different economic institutions lead to different distributions of resources. Consequently, there will typically be a conflict of interest among various groups and individuals over the choice of economic institutions. So how are equilibrium economic institutions determined? If there are, for example, two groups with opposing preferences over the set of economic institutions, which group’s preferences will prevail? The answer depends on the political power of the two groups. Although the efficiency of one set of economic institutions compared with another may play a role in this choice, political power will be the ultimate arbiter. Whichever group has more political power is likely to secure the set of economic institutions that it prefers. This leads to the second building block of our framework:

$$
\text{political power}_t \implies \text{economic institutions}_t
$$

3. Implicit in the notion that political power determines economic institutions is the idea that there are conflicting interests over the distribution of resources and therefore indirectly over the set of economic institutions. But why do the groups with conflicting interests not agree on the set of economic institutions that maximize aggregate growth (the size of the aggregate pie) and then use their political power simply to determine the distribution of the gains? Why does the exercise of political power lead to economic inefficiencies and even poverty? We will explain that this is because there are commitment problems inherent in the use of political power. Individuals who have political power
cannot commit not to use it in their best interests, and this commitment problem creates an inseparability between efficiency and distribution because credible compensating transfers and side-payments cannot be made to offset the distributional consequences of any particular set of economic institutions.

4. The distribution of political power in society is also endogenous, however. In our framework, it is useful to distinguish between two components of political power, which we refer to as *de jure* (institutional) and *de facto* political power. Here de jure political power refers to power that originates from the *political institutions* in society. Political institutions, similarly to economic institutions, determine the constraints on and the incentives of the key actors, but this time in the political sphere. Examples of political institutions include the form of government, for example, democracy vs. dictatorship or autocracy, and the extent of constraints on politicians and political elites. For example, in a monarchy, political institutions allocate all de jure political power to the monarch, and place few constraints on its exercise. A constitutional monarchy, in contrast, corresponds to a set of political institutions that reallocates some of the political power of the monarch to a parliament, thus effectively constraining the political power of the monarch. This discussion therefore implies that:

\[
\text{political institutions}_t \implies \text{de jure political power}_t
\]

5. There is more to political power than political institutions, however. A group of individuals, even if they are not allocated power by political institutions, for example as specified in the constitution, may nonetheless possess political power. Namely, they can revolt, use arms, hire mercenaries, co-opt the military, or use economically costly but largely peaceful protests in order to impose their wishes on society. We refer to this type of political power as *de facto* political power, which itself has two sources. First, it depends on the ability of the group in question to solve its collective action problem, i.e., to ensure that people act together, even when any individual may have an incentive to free ride. For example, peasants in the Middle Ages, who were given no political power by the constitution, could sometimes solve the collective action problem and undertake a revolt against the authorities. Second, the *de facto* power of a group depends on its economic resources, which determine both their ability to use (or misuse) existing political institutions and also their option to hire and use force against different groups. Since we do not yet have a satisfactory theory of when groups are able to solve their collective action problems, our focus will be on the second source of *de facto* political
power, hence:

\[ \text{distribution of resources}_t \implies \text{de facto political power}_t \]

6. This brings us to the evolution of one of the two main state variables in our framework, political institutions (the other state variable is the distribution of resources, including distribution of physical and human capital stocks etc.). Political institutions and the distribution of resources are the state variables in this dynamic system because they typically change relatively slowly, and more importantly, they determine economic institutions and economic performance both directly and indirectly. Their direct effect is straightforward to understand. If political institutions place all political power in the hands of a single individual or a small group, economic institutions that provide protection of property rights and equal opportunity for the rest of the population are difficult to sustain. The indirect effect works through the channels discussed above: political institutions determine the distribution of de jure political power, which in turn affects the choice of economic institutions. This framework therefore introduces a natural concept of a hierarchy of institutions, with political institutions influencing equilibrium economic institutions, which then determine economic outcomes.

Political institutions, though slow changing, are also endogenous. Societies transition from dictatorship to democracy, and change their constitutions to modify the constraints on power holders. Since, like economic institutions, political institutions are collective choices, the distribution of political power in society is the key determinant of their evolution. This creates a tendency for persistence: political institutions allocate de jure political power, and those who hold political power influence the evolution of political institutions, and they will generally opt to maintain the political institutions that give them political power. However, de facto political power occasionally creates changes in political institutions. While these changes are sometimes discontinuous, for example when an imbalance of power leads to a revolution or the threat of revolution leads to major reforms in political institutions, often they simply influence the way existing political institutions function, for example, whether the rules laid down in a particular constitution are respected as in most functioning democracies, or ignored as in current-day Zimbabwe. Summarizing this discussion, we have:

\[ \text{political power}_t \implies \text{political institutions}_{t+1} \]

Putting all these pieces together, a schematic (and simplistic) representation of our
framework is as follows:

\[
\begin{align*}
\text{political} & \quad \Rightarrow \quad \text{de jure political power} \\
\text{institutions}_t & \quad \& \quad \Rightarrow \quad \text{economic} \\
\text{distribution} & \quad \text{of resources}_t \quad \Rightarrow \quad \text{de facto political power}_t \\
& \quad \Rightarrow \quad \text{political institutions}_t \quad \Rightarrow \quad \text{economic performance}_t \\
& \quad \& \quad \text{distribution} \\
& \quad \quad \text{of resources}_{t+1}
\end{align*}
\]

The two state variables are political institutions and the distribution of resources, and the knowledge of these two variables at time \( t \) is sufficient to determine all the other variables in the system. While political institutions determine the distribution of de jure political power in society, the distribution of resources influences the distribution of de facto political power at time \( t \). These two sources of political power, in turn, affect the choice of economic institutions and influence the future evolution of political institutions. Economic institutions determine economic outcomes, including the aggregate growth rate of the economy and the distribution of resources at time \( t + 1 \). Although economic institutions are the essential factor shaping economic outcomes, they are themselves endogenous and determined by political institutions and distribution of resources in society.

There are two sources of persistence in the behavior of the system: first, political institutions are durable, and typically, a sufficiently large change in the distribution of political power is necessary to cause a change in political institutions, such as a transition from dictatorship to democracy. Second, when a particular group is rich relative to others, this will increase its de facto political power and enable it to push for economic and political institutions favorable to its interests. This will tend to reproduce the initial relative wealth disparity in the future. Despite these tendencies for persistence, the framework also emphasizes the potential for change. In particular, “shocks”, including changes in technologies and the international environment, that modify the balance of (de facto) political power in society and can lead to major changes in political institutions and therefore in economic institutions and economic growth.

A brief example might be useful to clarify these notions before commenting on some of the underlying assumptions and discussing comparative statics. Consider the development of property rights in Europe during the Middle Ages. There is no doubt that lack of property rights for landowners, merchants and proto-industrialists was detrimental to economic growth during this epoch. Since political institutions at the time placed political power in the hands of kings and various types of hereditary monarchies,
such rights were largely decided by these monarchs. Unfortunately for economic growth, while monarchs had every incentive to protect their own property rights, they did not generally enforce the property rights of others. On the contrary, monarchs often used their powers to expropriate producers, impose arbitrary taxation, renege on their debts, and allocate the productive resources of society to their allies in return for economic benefits or political support. Consequently, economic institutions during the Middle Ages provided little incentive to invest in land, physical or human capital, or technology, and failed to foster economic growth. These economic institutions also ensured that the monarchs controlled a large fraction of the economic resources in society, solidifying their political power and ensuring the continuation of the political regime.

The seventeenth century, however, witnessed major changes in the economic and political institutions that paved the way for the development of property rights and limits on monarchs’ power, especially in England after the Civil War of 1642 and the Glorious Revolution of 1688, and in the Netherlands after the Dutch Revolt against the Hapsburgs. How did these major institutional changes take place? In England, for example, until the sixteenth century the king also possessed a substantial amount of de facto political power, and leaving aside civil wars related to royal succession, no other social group could amass sufficient de facto political power to challenge the king. But changes in the English land market (Tawney, 1941) and the expansion of Atlantic trade in the sixteenth and seventeenth centuries (Acemoglu, Johnson and Robinson, 2002b) gradually increased the economic fortunes, and consequently the de facto power of landowners and merchants. These groups were diverse, but contained important elements that perceived themselves as having interests in conflict with those of the king: while the English kings were interested in predating against society to increase their tax incomes, the gentry and merchants were interested in strengthening their property rights.

By the seventeenth century, the growing prosperity of the merchants and the gentry, based both on internal and overseas, especially Atlantic, trade, enabled them to field military forces capable of defeating the king. This de facto power overcame the Stuart monarchs in the Civil War and Glorious Revolution, and led to a change in political institutions that stripped the king of much of his previous power over policy. These changes in the distribution of political power led to major changes in economic institutions, strengthening the property rights of both land and capital owners and spurred a process of financial and commercial expansion. The consequence was rapid economic growth, culminating in the Industrial Revolution, and a very different distribution of
economic resources from that in the Middle Ages.

It is worth returning at this point to two critical assumptions in our framework. First, why do the groups with conflicting interests not agree on the set of economic institutions that maximize aggregate growth? So in the case of the conflict between the monarchy and the merchants, why does the monarchy not set up secure property rights to encourage economic growth and tax some of the benefits? Second, why do groups with political power want to change political institutions in their favor? For instance, in the context of the example above, why did the gentry and merchants use their de facto political power to change political institutions rather than simply implement the policies they wanted? The answers to both questions revolve around issues of commitment and go to the heart of our framework.

The distribution of resources in society is an inherently conflictual, and therefore political, decision. As mentioned above, this leads to major commitment problems, since groups with political power cannot commit to not using their power to change the distribution of resources in their favor. For example, economic institutions that increased the security of property rights for land and capital owners during the Middle Ages would not have been credible as long as the monarch monopolized political power. He could promise to respect property rights, but then at some point, renege on his promise, as exemplified by the numerous financial defaults by medieval kings (e.g., Veitch, 1986). Credible secure property rights necessitated a reduction in the political power of the monarch. Although these more secure property rights would foster economic growth, they were not appealing to the monarchs who would lose their rents from predation and expropriation as well as various other privileges associated with their monopoly of political power. This is why the institutional changes in England as a result of the Glorious Revolution were not simply conceded by the Stuart kings. James II had to be deposed for the changes to take place.

The reason why political power is often used to change political institutions is related. In a dynamic world, individuals care not only about economic outcomes today but also in the future. In the example above, the gentry and merchants were interested in their profits and therefore in the security of their property rights, not only in the present but also in the future. Therefore, they would have liked to use their (de facto) political power to secure benefits in the future as well as the present. However, commitment to future allocations (or economic institutions) was not possible because decisions in the future would be decided by those who had political power in the future with little reference to past promises. If the gentry and merchants would have been sure to maintain their de
facto political power, this would not have been a problem. However, de facto political power is often transient, for example because the collective action problems that are solved to amass this power are likely to resurface in the future, or other groups, especially those controlling de jure power, can become stronger in the future. Therefore, any change in policies and economic institutions that relies purely on de facto political power is likely to be reversed in the future. In addition, many revolutions are followed by conflict within the revolutionaries. Recognizing this, the English gentry and merchants strove not just to change economic institutions in their favor following their victories against the Stuart monarchy, but also to alter political institutions and the future allocation of de jure power. Using political power to change political institutions then emerges as a useful strategy to make gains more durable. The framework that we propose, therefore, emphasizes the importance of political institutions, and changes in political institutions, as a way of manipulating future political power, and thus indirectly shaping future, as well as present, economic institutions and outcomes.

This framework, though abstract and highly simple, enables us to provide some preliminary answers to our main question: why do some societies choose “good economic institutions”? At this point, we need to be more specific about what good economic institutions are. A danger we would like to avoid is that we define good economic institutions as those that generate economic growth, potentially leading to a tautology. This danger arises because a given set of economic institutions may be relatively good during some periods and bad during others. For example, a set of economic institutions that protects the property rights of a small elite might not be inimical to economic growth when all major investment opportunities are in the hands of this elite, but could be very harmful when investments and participation by other groups are important for economic growth (see Acemoglu, 2003b). To avoid such a tautology and to simplify and focus the discussion, throughout we think of good economic institutions as those that provide security of property rights and relatively equal access to economic resources to a broad cross-section of society. Although this definition is far from requiring equality of opportunity in society, it implies that societies where only a very small fraction of the population have well-enforced property rights do not have good economic institutions. Consequently, as we will see in some of the historical cases discussed below, a given set of economic institutions may have very different implications for economic growth depending on the technological possibilities and opportunities.

Given this definition of good economic institutions as providing secure property rights for a broad cross-section of society, our framework leads to a number of important com-
parative statics, and thus to an answer to our basic question. First, political institutions that place checks on those who hold political power, for example, by creating a balance of power in society, are useful for the emergence of good economic institutions. This result is intuitive; without checks on political power, power holders are more likely to opt for a set of economic institutions that are beneficial for themselves and detrimental for the rest of society, which will typically fail to protect property rights of a broad cross-section of people. Second, good economic institutions are more likely to arise when political power is in the hands of a relatively broad group with significant investment opportunities. The reason for this result is that, everything else equal, in this case power holders will themselves benefit from secure property rights. Third, good economic institutions are more likely to arise and persist when there are only limited rents that power holders can extract from the rest of society, since such rents would encourage them to opt for a set of economic institutions that make the expropriation of others possible. These comparative statics therefore place political institutions at the center of the story, as emphasized by our term “hierarchy of institutions” above. Political institutions are essential both because they determine the constraints on the use of (de facto and de jure) political power and also which groups hold de jure political power in society. We will see below how these comparative statics help us understand institutional differences across countries and over time in a number of important historical examples.

1.3 Outline

In the next section we discuss how economic institutions constitute the basis for a fundamental theory of growth, and we contrast this with other potential fundamental theories. In section 3 we consider some empirical evidence that suggests a key role for economic institutions in determining long-run growth. We also emphasize some of the key problems involved in establishing a causal relationship between economic institutions and growth. We then show in section 4 how the experience of European colonialism can be used as a ‘natural experiment’ which can address these problems. Having established the central causal role of economic institutions and their importance relative to other factors in cross-country differences in economic performance, the rest of the paper focuses on developing a theory of economic institutions. Section 5 discusses four types of explanation for why countries have different institutions, and argues that

---

2 The reason why we inserted the caveat of “a relatively broad group” is that when a small group with significant investment opportunities holds power, they may sometimes opt for an oligarchic system where their own property rights are protected, but those of others are not (see Acemoglu, 2003b).
the most plausible is the social conflict view. According to this theory, bad institutions arise because the groups with political power benefit from bad institutions. The emphasis on social conflict arises naturally from our observation above that economic institutions influence the distribution of resources as well as efficiency. Different groups or individuals will therefore prefer different institutions and conflict will arise as each tries to get their own way. Section 6 delves deeper into questions of efficiency and asks why a political version of the Coase Theorem does not hold. We emphasize the idea that commitment problems are intrinsic to the exercise of political power. In section 7 we argue that a series of historical examples of diverging economic institutions are best explained by the social conflict view. These examples illustrate how economic institutions are determined by the distribution of political power, and how this distribution is influenced by political institutions. Section 8 puts these ideas together to build our theory of institutions. In section 9 we then consider two more extended examples of the theory in action, the rise of constitutional rule in early modern Europe, and the creation of mass democracy, particularly in Britain, in the nineteenth and twentieth centuries. Section 10 concludes with a discussion of where this research program can go next.

2 Fundamental Causes of Income Differences

We begin by taking a step back. The presumption in the introduction was that economic institutions matter, and should in fact be thought of as one of the key fundamental causes of economic growth and cross-country differences in economic performance. How do we know this?

2.1 Three Fundamental Causes

If standard economic models of factor accumulation and endogenous technical change only provide proximate explanations of comparative growth, what types of explanations would constitute fundamental ones? Though there is no conventional wisdom on this, we can distinguish three such theories: the first set of theories, our main focus in this chapter, emphasize the importance of economic institutions, which influence economic outcomes by shaping economic incentives; the second emphasize geography, and the third emphasize the importance of culture (a fourth possibility is that differences are due to “luck,” some societies were just lucky; however we do not believe that differences in luck by themselves constitute a sufficient fundamental causes of cross-country income differences).
2.1.1 Economic Institutions

At its core, the hypothesis that differences in economic institutions are the fundamental cause of different patterns of economic growth is based on the notion that it is the way that humans themselves decide to organize their societies that determines whether or not they prosper. Some ways of organizing societies encourage people to innovate, to take risks, to save for the future, to find better ways of doing things, to learn and educate themselves, solve problems of collective action and provide public goods. Others do not.

The idea that the prosperity of a society depends on its economic institutions goes back at least to Adam Smith, for example in his discussions of mercantilism and the role of markets, and was prominent in the work of many nineteenth century scholars such as John Stuart Mill (see the discussion in Jones, 1981): societies are economically successful when they have ‘good’ economic institutions and it is these institutions that are the cause of prosperity. We can think of these good economic institutions as consisting of an inter-related cluster of things. There must be enforcement of property rights for a broad cross-section of society so that all individuals have an incentive to invest, innovate and take part in economic activity. There must also be some degree of equality of opportunity in society, including such things as equality before the law, so that those with good investment opportunities can take advantage of them.³

One could think of other types of economic institutions, for instance markets. Traditional accounts of economic growth by historians, following the lead of Adam Smith, emphasized the spread of markets (Pirenne, 1937, Hicks, 1969) and more recent theories of comparative development are also based on differences in various economic institutions. Models of poverty traps in the tradition of Rosenstein-Rodan (1943), Murphy, Vishny and Shleifer (1989a,b) and Acemoglu (1995, 1997), are based on the idea that market imperfections can lead to the existence of multiple Pareto-ranked equilibria. As a consequence a country can get stuck in a Pareto inferior equilibrium, associated with poverty, but getting out of such a trap necessitates coordinated activities that the market cannot deliver. The literature initiated by Banerjee and Newman (1993) and Galor and Zeira (1993) is based on the idea that when capital markets are imperfect, the distribution of wealth matters for who can invest and societies with skewed income distributions can be stuck in poverty.

³In Acemoglu, Johnson and Robinson (2001), we coined the term institutions of private property for a cluster of would economic institutions, including the rule of law and the enforcement of property rights, and the term extractive institutions to designate institutions under which the rule of law and property rights are absent for large majorities of the population.
These theories provide interesting models of how incentives depend on expectations of others’ behavior or the distribution of wealth given an underlying set of market imperfections. They take the market structure largely as given, however. We believe that the structure of markets is endogenous, and partly determined by property rights. Once individuals have secure property rights and there is equality of opportunity, the incentives will exist to create and improve markets (even though achieving perfect markets would be typically impossible). Thus we expect differences in markets to be an outcome of differing systems of property rights and political institutions, not unalterable characteristics responsible for cross-country differences in economic performance. This motivates our focus on economic institutions related to the enforcement of the property rights of a broad cross-section of society.

2.1.2 Geography

While institutional theories emphasize the importance of man-made factors shaping incentives, an alternative is to focus on the role of “nature”, that is, on the physical and geographical environment. In the context of understanding cross-country differences in economic performance, this approach emphasizes differences in geography, climate and ecology that determine both the preferences and the opportunity set of individual economic agents in different societies. We refer to this broad approach as the “geography hypothesis”. There are at least three main versions of the geography hypothesis, each emphasizing a different mechanism for how geography affects prosperity.

First, climate may be an important determinant of work effort, incentives, or even productivity. This idea dates back at least to the famous French philosopher, Montesquieu ([1748], 1989), who wrote in his classic book The Spirit of the Laws: “The heat of the climate can be so excessive that the body there will be absolutely without strength. So, prostration will pass even to the spirit; no curiosity, no noble enterprise, no generous sentiment; inclinations will all be passive there; laziness there will be happiness,” and “People are ... more vigorous in cold climates. The inhabitants of warm countries are, like old men, timorous; the people in cold countries are, like young men, brave”. One of the founders of modern economics Marshall is another prominent figure who emphasized the importance of climate, arguing: “vigor depends partly on race qualities: but these, so far as they can be explained at all, seem to be chiefly due to climate” (1890, p. 195).

Second, geography may determine the technology available to a society, especially in agriculture. This view is developed by an early Nobel Prize winner in economics,
Myrdal, who wrote “serious study of the problems of underdevelopment ... should take into account the climate and its impacts on soil, vegetation, animals, humans and physical assets—in short, on living conditions in economic development” (1968, volume 3, p. 2121). More recently, Diamond espouses this view, “... proximate factors behind Europe’s conquest of the Americas were the differences in all aspects of technology. These differences stemmed ultimately from Eurasia’s much longer history of densely populated ... [societies dependent on food production],” which was in turn determined by geographical differences between Europe and the Americas (1997, p. 358). The economist Sachs has been a recent and forceful proponent of the importance of geography in agricultural productivity, stating that “By the start of the era of modern economic growth, if not much earlier, temperate-zone technologies were more productive than tropical-zone technologies ...” (2001, p. 2).

The third variant of the geography hypothesis, especially popular over the past decade, links poverty in many areas of the world to their “disease burden,” emphasizing that: “The burden of infectious disease is similarly higher in the tropics than in the temperate zones” (Sachs, 2000, p. 32). Bloom and Sachs (1998) claim that the prevalence of malaria, a disease which kills millions of children every year in sub-Saharan Africa, reduces the annual growth rate of sub-Saharan African economies by more than 1.3 percent a year (this is a large effect, implying that had malaria been eradicated in 1950, income per capita in sub-Saharan Africa would be double of what it is today).

2.1.3 Culture

The final fundamental explanation for economic growth emphasizes the idea that different societies (or perhaps different races or ethnic groups) have different cultures, because of different shared experiences or different religions. Culture is viewed as a key determinant of the values, preferences and beliefs of individuals and societies and, the argument goes, these differences play a key role in shaping economic performance.

At some level, culture can be thought to influence equilibrium outcomes for a given set of institutions. Possibly there are multiple equilibria connected with any set of institutions and differences in culture mean that different societies will coordinate on different equilibria. Alternatively, as argued by Greif (1993), different cultures generate different sets of beliefs about how people behave and this can alter the set of equilibria for a given specification of institutions (for example, some beliefs will allow punishment strategies to be used whereas others will not).
The most famous link between culture and economic development is that proposed by Weber (1930) who argued that the origins of industrialization in western Europe could be traced to the Protestant reformation and particularly the rise of Calvinism. In his view, the set of beliefs about the world that was intrinsic to Protestantism were crucial to the development of capitalism. Protestantism emphasized the idea of predestination in the sense that some individuals were ‘chosen’ while others were not. “We know that a part of humanity is saved, the rest damned. To assume that human merit or guilt play a part in determining this destiny would be to think of God’s absolutely free decrees, which have been settled from eternity, as subject to change by human influence, an impossible contradiction” (Weber, 1930, p. 60).

But who had been chosen and who not? Calvin did not explain this. Weber (1930, p. 66) notes “Quite naturally this attitude was impossible for his followers ... for the broad mass of ordinary men ... So wherever the doctrine of predestination was held, the question could not be suppressed whether there was any infallible criteria by which membership of the elect could be known.” Practical solutions to this problem were quickly developed, “... in order to attain that self-confidence intense worldly activity is recommended as the most suitable means. It and it alone disperses religious doubts and gives the certainly of grace” Weber (1930, pp. 66-67).

Thus “however useless good works might be as a means of attaining salvation ... nevertheless, they are indispensable as a sign of election. They are the technical means, not of purchasing salvation, but of getting rid of the fear of damnation” (p. 69). Though economic activity was encouraged, enjoying the fruits of such activity was not. “Waste of time is ... the first and in principle the deadliest of sins. The span of human life is infinitely short and precious to make sure of one’s own election. Loss of time through sociability, idle talk, luxury, even more sleep than is necessary for health ... is worthy of absolute moral condemnation ... Unwillingness to work is symptomatic of the lack of grace” (pp. 104-105).

Thus Protestantism led to a set of beliefs which emphasized hard work, thrift, saving, and where economic success was interpreted as consistent with (if not actually signalling) being chosen by God. Weber contrasted these characteristics of Protestantism with those of other religions, such as Catholicism, which he argued did not promote capitalism. For instance on his book on Indian religion he argued that the caste system blocked capitalist development (Weber, 1958, p. 112).

More recently, scholars, such as Landes (1998), have also argued that the origins of Western economic dominance are due to a particular set of beliefs about the world and
how it could be transformed by human endeavor, which is again linked to religious differences. Although Barro and McCleary (2003) provide evidence of a positive correlation between the prevalence of religious beliefs, notably about hell and heaven, and economic growth, this evidence does not show a causal effect of religion on economic growth, since religious beliefs are endogenous both to economic outcomes and to other fundamental causes of income differences (points made by Tawney, 1926, and Hill, 1961b, in the context of Weber’s thesis).

Ideas about how culture may influence growth are not restricted to the role of religion. Within the literature trying to explain comparative development there have been arguments that there is something special about particular cultural endowments, usually linked to particular nation states. For instance, Latin America may be poor because of its Iberian heritage, while North America is prosperous because of its Anglo-Saxon heritage (Vélez, 1994). In addition, a large literature in anthropology argues that societies may become ‘dysfunctional’ or ‘maladapted’ in the sense that they adopt a system of beliefs or ways of operating which do not promote the success or prosperity of the society (see Edgerton, 1992, for a survey of this literature). The most famous version of such an argument is due to Banfield (1958) who argued that the poverty of Southern Italy was due to the fact that people had adopted a culture of “amoral familiarism” where they only trusted individuals of their own families and refused to cooperate or trust anyone else. This argument was revived in the extensive empirical study of Putnam (1993) who characterized such societies as lacking “social capital”. Although Putnam and others, for example, Knack and Keefer (1997) and Durlauf and Fafchamps (2003), document positive correlations between measures of social capital and various economic outcomes, there is no evidence of a causal effect, since, as with religious beliefs discussed above, measures of social capital are potentially endogenous.

3 Institutions Matter

We now argue that there is convincing empirical support for the hypothesis that differences in economic institutions, rather than geography or culture, cause differences in incomes per-capita. Consider first Figure 1.

This shows the cross-country bivariate relationship between the log of GDP per-capita in 1995 and a broad measure of property rights, “protection against expropriation risk”, averaged over the period 1985 to 1995. The data on economic institutions come from Political Risk Services, a private company which assesses the risk that investments will be expropriated in different countries. These data, first used by Knack and Keefer
(1995) and subsequently by Hall and Jones (1999) and Acemoglu, Johnson and Robinson (2001, 2002a) are imperfect as a measure of economic institutions, but the findings are robust to using other available measures of economic institutions. The scatter plot shows that countries with more secure property rights, i.e., better economic institutions, have higher average incomes.

It is tempting to interpret Figure 1 as depicting a causal relationship (i.e., as establishing that secure property rights cause prosperity). Nevertheless, there are well known problems with making such an inference. First, there could be reverse causation – perhaps only countries that are sufficiently wealthy can afford to enforce property rights. More importantly, there might be a problem of omitted variable bias. It could be something else, e.g., geography, that explains both why countries are poor and why they have insecure property rights. Thus if omitted factors determine institutions and incomes, we would spuriously infer the existence of a causal relationship between economic institutions and incomes when in fact no such relationship exists. Trying to estimate the relationship between institutions and prosperity using Ordinary Least Squares, as was done by Knack and Keefer (1995) and Barro (1997) could therefore result in biased regression coefficients.

To further illustrate these potential identification problems, suppose that climate, or geography more generally, matters for economic performance. In fact, a simple scatterplot shows a positive association between latitude (the absolute value of distance from the equator) and income per capita. Montesquiou, however, not only claimed that warm climate makes people lazy and thus unproductive, but also unfit to be governed by democracy. He argued that despotism would be the political system in warm climates. Therefore, a potential explanation for the patterns we see in Figure 1 is that there is an omitted factor, geography, which explains both economic institutions and economic performance. Ignoring this potential third factor would lead to mistaken conclusions.

Even if Montesquiou’s story appears both unrealistic and condescending to our modern sensibilities, the general point should be taken seriously: the relationship shown in Figure 1, and for that matter that shown in Figure 2, is not causal. As we pointed out in the context of the effect of religion or social capital on economic performance, these types of scatterplots, correlations, or their multidimensional version in OLS regressions, cannot establish causality.

What can we do? The solution to these problems of inference is familiar in microeconometrics: find a source of variation in economic institutions that should have no effect on economic outcomes, or depending on the context, look for a natural experiment.
As an example, consider first one of the clearest natural experiments for institutions.

3.1 The Korean Experiment

Until the end of World War II, Korea was under Japanese occupation. Korean independence came shortly after the Japanese Emperor Hirohito announced the Japanese surrender on August 15, 1945. After this date, Soviet forces entered Manchuria and North Korea and took over the control of these provinces from the Japanese. The major fear of the United States during this time period was the takeover of the entire Korean peninsular either by the Soviet Union or by communist forces under the control of the former guerrilla fighter, Kim Il Sung. U.S. authorities therefore supported the influential nationalist leader Syngman Rhee, who was in favor of separation rather than a united communist Korea. Elections in the South were held in May 1948, amidst a widespread boycott by Koreans opposed to separation. The newly elected representatives proceeded to draft a new constitution and established the Republic of Korea to the south of the 38th parallel. The North became the Democratic People’s Republic of Korea, under the control of Kim Il Sung. These two independent countries organized themselves in very different ways and adopted completely different sets of institutions. The North followed the model of Soviet socialism and the Chinese Revolution in abolishing private property of land and capital. Economic decisions were not mediated by the market, but by the communist state. The South instead maintained a system of private property and the government, especially after the rise to power of Park Chung Hee in 1961, attempted to use markets and private incentives in order to develop the economy.

Before this “natural experiment” in institutional change, North and South Korea shared the same history and cultural roots. In fact, Korea exhibited an unparalleled degree of ethnic, linguistic, cultural, geographic and economic homogeneity. There are few geographic distinctions between the North and South, and both share the same disease environment. For example, the CIA Factbook describes the climate of North Korea as “temperate with rainfall concentrated in summer” and that of South Korea as “temperate, with rainfall heavier in summer than winter”. In terms of terrain North Korea is characterized as consisting of “mostly hills and mountains separated by deep, narrow valleys; coastal plains wide in west, discontinuous in east,” while South Korea is “mostly hills and mountains; wide coastal plains in west and south”. In terms of natural resources North Korea is better endowed with significant reserves of coal, lead, tungsten, zinc, graphite, magnesite, iron ore, copper, gold, pyrites, salt, fluor spar, hydropower. South Korea’s natural resources are “coal, tungsten, graphite, molybdenum,
lead, hydropower potential.” Both countries share the same geographic possibilities in terms of access to markets and the cost of transportation.

Other man-made initial economic conditions were also similar, and if anything, advantaged the North. For example, there was significant industrialization during the colonial period with the expansion of both Japanese and indigenous firms. Yet this development was concentrated more in the North than the South. For instance, the large Japanese zaibatsu of Noguchi, which accounted for one third of Japanese investment in Korea, was centered in the North. It built large hydroelectric plants, including the Suiho dam on the Yalu river, second in the world only to the Boulder dam on the Colorado river. It also created Nippon Chisso, the second largest chemical complex in the world that was taken over by the North Korean state. Finally, in Ch’ongjin North Korea also had the largest port on the Sea of Japan. All in all, despite some potential advantages for the North, Maddison (2001) estimates that at the time of separation, North and South Korea had approximately the same income per capita.

We can therefore think of the splitting on the Koreas 50 years ago as a natural experiment that we can use to identify the causal influence of a particular dimension of institutions on prosperity. Korea was split into two, with the two halves organized in radically different ways, and with geography, culture and many other potential determinants of economic prosperity held fixed. Thus any differences in economic performance can plausibly be attributed to differences in institutions.

Consistent with the hypothesis that it is institutional differences that drive comparative development, since separation, the two Koreas have experienced dramatically diverging paths of economic development: see Figure 3. By the late 1960’s South Korea was transformed into one of the Asian “miracle” economies, experiencing one of the most rapid surges of economic prosperity in history while North Korea stagnated. By 2000 the level of income in South Korea was $16,100 while in North Korea it was only $1,000. By 2000 the South had become a member of the Organization of Economic Cooperation and Development, the rich nations club, while the North had a level of per-capita income about the same as a typical sub-Saharan African country. There is only one plausible explanation for the radically different economic experiences on the two Koreas after 1950: their very different institutions led to divergent economic outcomes. In this context, it is noteworthy that the two Koreas not only shared the same geography, but also the same culture.

4Such initial differences were probably eradicated by the intensive bombing campaign that the United States unleashed in the early 1950’s on North Korea (see Cumings, 2004, chapter 1).
It is possible that Kim Il Sung and Communist Party members in the North believed that communist policies would be better for the country and the economy in the late 1940s. However, by the 1980s it was clear that the communist economic policies in the North were not working. The continued efforts of the leadership to cling to these policies and to power can only be explained by those leaders wishing to look after their own interests at the expense of the population at large. Bad institutions are therefore kept in place, clearly not for the benefit of society as a whole, but for the benefit of the ruling elite, and this is a pattern we encounter in most cases of institutional failure that we discuss in detail below.

However convincing on its own terms, the evidence from this natural experiment is not sufficient for the purposes of establishing the importance of economic institutions as the primary factor shaping cross-country differences in economic prosperity. First, this is only one case, and in the better-controlled experiments in the natural sciences, a relatively large sample is essential. Second, here we have an example of an extreme case, the difference between a market-oriented economy and a communist one. Few social scientists today would deny that a lengthy period of totalitarian centrally planned rule has significant economic costs. And yet, many might argue that differences in economic institutions among capitalist economies or among democracies are not the major factor leading to differences in their economic trajectories. To establish the major role of economic institutions in the prosperity and poverty of nations we need to look at a larger scale “natural experiment” in institutional divergence.

3.2 The Colonial Experiment

The colonization of much of the world by Europeans provides such a large scale natural experiment. Beginning in the early fifteenth century and massively intensifying after 1492, Europeans conquered many other nations. The colonization experience transformed the institutions in many diverse lands conquered or controlled by Europeans. Most importantly, Europeans imposed very different sets of institutions in different parts of their global empire, as exemplified most sharply by the contrast to the economic institutions in the northeast of America to those in the plantation societies of the Caribbean. As a result, while geography was held constant, Europeans initiated large changes in economic institutions, in the social organization of different societies. We will now show that this experience provides evidence which conclusively establishes the central role of economic institutions in development. Given the importance of this material and the details we need to provide, we discuss the colonial experience in the
next section.

4 The Reversal of Fortune

The impact of European colonialism on economic institutions is perhaps most dramatically conveyed by a single fact—historical evidence shows that there has been a remarkable Reversal of Fortune in economic prosperity within former European colonies. Societies like the Mughals in India, and the Aztecs and the Incas in the Americas were among the richest civilizations in 1500, yet the nation states that now coincide with the boundaries of these empires are among the poorer societies of today. In contrast, countries occupying the territories of the less-developed civilizations in North America, New Zealand and Australia are now much richer than those in the lands of the Mughals, Aztecs and Incas.

4.1 The reversal among the former colonies

The Reversal of Fortune is not confined to such comparisons. Using reasonable proxies for prosperity before modern times, we can show that it is a much more systematic phenomenon. Our proxies for income per capita in pre-industrial societies are urbanization rates and population density. Only societies with a certain level of productivity in agriculture and a relatively developed system of transport and commerce can sustain large urban centers and a dense population. Figure 4 shows the relationship between income per capita and urbanization (fraction of the population living in urban centers with greater than 5,000 inhabitants) today, and demonstrates that in the current period there is a significant relationship between urbanization and prosperity.

Naturally, high rates of urbanization do not mean that the majority of the population lived in prosperity. In fact, before the twentieth century urban centers were often centers of poverty and ill health. Nevertheless, urbanization is a good proxy for average income per capita in society, which closely corresponds to the measure we are using to look at prosperity.

Figures 5 and 6 show the relationship between income per capita today and urbanization rates and (log) population density in 1500 for the sample of European colonies. We pick 1500 since it is before European colonization had an effect on any of these

---

5 The sample includes the countries colonized by the Europeans between the 15th and the 19th centuries as part of their overseas expansion after the discovery of the New World and the rounding of the Cape of Good Hope. It therefore excludes Ireland, parts of the Russian Empire and also the Middle East and countries briefly controlled by European powers as U.N. Mondays during the 20th century.
societies. A strong negative relationship, indicating a reversal in the rankings in terms of economic prosperity between 1500 and today, is clear in both figures. In fact, the figures show that in 1500 the temperate areas were generally less prosperous than the tropical areas, but this pattern too was reversed by the twentieth century.

The urbanization data for these Figures come from Bairoch (1988), Bairoch, Batou and Chèvre (1988), Chandler (1987), and Eggimann (1999). The data on population density are from McEvedy and Jones (1978). Details and further results are in Acemoglu, Johnson and Robinson (2002a).

There is something extraordinary about this reversal. For example, after the initial spread of agriculture there was remarkable persistence in urbanization and population density for all countries, including those which were to be subsequently colonized by Europeans. In Figures 7 and 8 we show the relationships for urbanization plotting separately the relationship between urbanization in 1000 and in 1500 for the samples of colonies and all other countries. Both figures show persistence, not reversal. Although Ancient Egypt, Athens, Rome, Carthage and other empires rose and fell, what these pictures show is that there was remarkable persistence in the prosperity of regions.

Moreover, reversal was not the general pattern in the world after 1500. Figure 9 shows that within countries not colonized by Europeans in the early modern and modern period, there was no reversal between 1500 and 1995. There is therefore no reason to think that what is going on in Figures 5 and 6 is some sort of natural reversion to the mean.

4.2 Timing of the Reversal

When did the reversal occur? One possibility is that it arose shortly after the conquest of societies by Europeans but Figures 10 and 11 show that the previously-poor colonies surpassed the former highly-urbanized colonies starting in the late eighteenth and early nineteenth centuries, and this went hand in hand with industrialization. Figure 10 shows average urbanization in colonies with relatively low and high urbanization in 1500. The initially high-urbanization countries have higher levels of urbanization and prosperity until around 1800. At that time the initially low-urbanization countries start to grow much more rapidly and a prolonged period of divergence begins. Figure 11 shows industrial production per capita in a number of countries. Although not easy to see in the figure, there was more industry (per capita and total) in India in 1750 than in the United States. By 1860, the United States and British colonies with relatively good economic institutions, such as Australia and New Zealand, began to move ahead
rapidly, and by 1953, a huge gap had opened up.

4.3 Interpreting the Reversal

Which of the three broad hypotheses about the sources of cross-country income differences are consistent with the reversal and its timing? These patterns are clearly inconsistent with simple geography based views of relative prosperity. In 1500 it was the countries in the tropics which were relatively prosperous, in 2003 it is the reverse. This makes it implausible to base a theory of relative prosperity today, as Sachs (2000, 2001) does, on the intrinsic poverty of the tropics. This argument is inconsistent with the historical evidence.

Nevertheless, following Diamond (1997), one could propose what Acemoglu, Johnson and Robinson (2002a) call a “sophisticated geography hypothesis” which claims that geography matters but in a time varying way. For example, Europeans created “latitude specific” technology, such as heavy metal ploughs, that only worked in temperate latitudes and not with tropical soils. Thus when Europe conquered most of the world after 1492, they introduced specific technologies that functioned in some places (the United States, Argentina, Australia) but not others (Peru, Mexico, West Africa). However, the timing of the reversal, coming as it does in the nineteenth century, is inconsistent with the most natural types of sophisticated geography hypotheses. Europeans may have had latitude specific technologies, but the timing implies that these technologies must have been industrial, not agricultural, and it is difficult to see why industrial technologies do not function in the tropics (and in fact, they have functioned quite successfully in tropical Singapore and Hong Kong).6

Similar considerations weigh against the culture hypothesis. Although culture is slow-changing the colonial experiment was sufficiently radical to have caused major changes in the cultures of many countries that fell under European rule. In addition, the destruction of many indigenous populations and immigration from Europe are likely to have created new cultures or at least modified existing cultures in major ways (see Vargas Llosa, 1989, for a fictionalized account of just such a cultural change). Nevertheless, the culture hypothesis does not provide a natural explanation for the reversal, and has nothing to say on the timing of the reversal. Moreover, we discuss below how econometric models that control for the effect of institutions on income do not find any evidence of an effect

---

6 A possible link is that proposed by Lewis (1978) who argued that tropical agriculture is less productive than temperate agriculture, and that an ‘agricultural revolution’ is a prerequisite to an industrial revolution because high agricultural productivity is needed to stimulate the demand for industrial goods.
of religion or culture on prosperity.

The most natural explanation for the reversal comes from the institutions hypothesis, which we discuss next.

4.4 Economic Institutions and the Reversal

Is the Reversal of Fortune consistent with a dominant role for economic institutions in comparative development? The answer is yes. In fact, once we recognize the variation in economic institutions created by colonization, we see that the Reversal of Fortune is exactly what the institutions hypothesis predicts.

In Acemoglu, Johnson and Robinson (2002a) we tested the connection between initial population density, urbanization, and the creation of good economic institutions. We showed that, others things equal, the higher the initial population density or the greater initial urbanization, the worse were subsequent institutions, including both institutions right after independence and today. Figures 12 and 13 show these relationships using the same measure of current economic institutions used in Figure 1, protection against expropriation risk today. They document that the relatively densely settled and highly urbanized colonies ended up with worse (or ‘extractive’) institutions, while sparsely-settled and non-urbanized areas received an influx of European migrants and developed institutions protecting the property rights of a broad cross-section of society. European colonialism therefore led to an institutional reversal, in the sense that the previously-richer and more-densely settled places ended up with worse institutions.

To be fair, it is possible that the Europeans did not actively introduce institutions discouraging economic progress in many of these places, but inherited them from previous civilizations there. The structure of the Mughal, Aztec and Inca empires were already very hierarchical with power concentrated in the hands of narrowly based ruling elites and structured to extract resources from the majority for the benefit of a minority. Often Europeans simply took over these existing institutions. Whether this is so is secondary for our focus, however. What matters is that in densely-settled and relatively-developed places it was in the interests of Europeans to have institutions facilitating the extraction of resources thus not respecting the property rights of the majority, while in the sparsely-settled areas it was in their interests to develop institutions protecting

---

7 The institutional reversal does not mean that institutions were necessarily better in the previously more densely-settled areas (see the next paragraph). It only implies a tendency for the relatively poorer and less densely-settled areas to end up with better institutions than previously-rich and more densely-settled areas.
property rights. These incentives led to an institutional reversal.

The institutional reversal, combined with the institutions hypothesis, predicts the Reversal of Fortune: relatively rich places got relatively worse economic institutions, and if these institutions are important, we should see them become relatively poor over time. This is exactly what we find with the Reversal of Fortune.

Moreover, the institutions hypothesis is consistent with the timing of the reversal. Recall that the institutions hypothesis links incentives to invest in physical and human capital and in technology to economic institutions, and argues that economic prosperity results from these investments. Therefore, economic institutions should become more important when there are major new investment opportunities. The opportunity to industrialize was the major investment opportunity of the nineteenth century. Countries that are rich today, both among the former European colonies and other countries, are those that industrialized successfully during this critical period.

4.5 Understanding the colonial experience

The explanation for the reversal that emerges from our discussion so far is one in which the economic institutions in various colonies were shaped by Europeans to benefit themselves. Moreover, because conditions and endowments differed between colonies, Europeans consciously created different economic institutions, which persisted and continue to shape economic performance. Why did Europeans introduce better institutions in previously-poor and unsettled areas than in previously-rich and densely-settled areas? The answer to this question relates to the comparative statics of our theoretical framework. Leaving a full discussion to later, we can note a couple of obvious ideas.

Europeans were more likely to introduce or maintain economic institutions facilitating the extraction of resources in areas where they would benefit from the extraction of resources. This typically meant areas controlled by a small group of Europeans, and areas offering resources to be extracted. These resources included gold and silver, valuable agricultural commodities such as sugar, but most importantly people. In places with a large indigenous population, Europeans could exploit the population, be it in the form of taxes, tributes or employment as forced labor in mines or plantations. This type of colonization was incompatible with institutions providing economic or civil rights to the majority of the population. Consequently, a more developed civilization and a denser population structure made it more profitable for the Europeans to introduce worse economic institutions.

In contrast, in places with little to extract, and in sparsely-settled places where the
Europeans themselves became the majority of the population, it was in their interests to introduce economic institutions protecting their own property rights.

4.6 Settlements, Mortality and Development

The initial conditions we have emphasized so far refer to indigenous population density and urbanization. In addition, the disease environments differed markedly among the colonies, with obvious consequences on the attractiveness of European settlement. As we noted above, when Europeans settled, they established institutions that they themselves had to live under. Therefore, whether Europeans could settle or not had an exogenous effect on the subsequent path of institutional development. In other words, if the disease environment 200 or more years ago affects outcomes today only through its effect on institutions today, then we can use this historical disease environment as an exogenous source of variation in current institutions. From an econometric point of view we have a valid instrument which will enable us to pin down the casual effect of economic institutions on prosperity.\(^8\)

We developed this argument in Acemoglu, Johnson and Robinson (2001) and investigated it empirically. We used initial conditions in the European colonies, particularly data from Curtin (1989, 1998) and Gutierrez (1986) on the mortality rates faced by Europeans (primarily soldiers, sailors, and bishops), as instruments for current economic institutions. The justification for this is that, outside of its effect on economic institutions during the colonial period, historical European mortality has no impact on current income levels. Figures 14 and 15 give scatter plots of this data against contemporaneous economic institutions and GDP per-capita. The sample is countries which were colonized by Europeans in the early modern and modern periods and thus excludes, among others, China, Japan, Korea, Thailand.

Figure 14 shows the very strong relationship between the historical mortality risk faced by Europeans and the current extent to which property rights are enforced. A bivariate regression has an \(R^2\) of 0.26. It also shows that there were very large differences in European mortality. Countries such as Australia, New Zealand and the United States were very healthy with life expectancy typically greater than in Britain. On the other hand mortality was extremely high in Africa, India and South-East Asia. Differential

\(^8\)Although European mortality is potentially correlated with indigenous mortality, which may determine income today, in practice local populations have developed much greater immunity to malaria and yellow fever. Thus the historical experience of European mortality is a valid instrument for institutional development. See Acemoglu, Johnson and Robinson (2001).
mortality was largely due to tropical diseases such as malaria and yellow fever and at the time it was not understood how these diseases arose nor how they could be prevented or cured.

In Acemoglu, Johnson and Robinson (2001) we showed, using European mortality as an instrument for the current enforcement of property rights, that most of the gap between rich and poor countries today is due to differences in economic institutions. More precisely, we showed (p. 1387) that if one took two typical—in the sense that they both lie on the regression line—countries with high and low expropriation risk, like Nigeria and Chile, then almost the entire difference in incomes per-capita between them could be explained by the differences in the security of property rights. We also presented regression evidence that showed that once the effect of economic institutions on GDP per-capita was properly controlled for, geographical variables, such as latitude, whether or not a country is land-locked and the current disease environment, have no explanatory power for current prosperity.

These ideas and results provide an interpretation of why there are strong correlations between geographical variables such as latitude and income per-capita. Basically this is because Europeans did not have immunity to tropical diseases during the colonial period and thus settler colonies tended, other things equal, to be created in temperate latitudes. Thus the historical creation of economic institutions was correlated with latitude. Without considering the role of economic institutions it is easy to find a spurious relationship between latitude and income per-capita. However, once economic institutions are properly controlled for, these relationships go away. There is no causal effect of geography on prosperity today, though geography may have been important historically in shaping economic institutions.

What about the role of culture? On the face of it, the Reversal of Fortune is consistent with cultural explanations of comparative growth. The Europeans not only brought new institutions, they also brought their own cultures. There seem to be three main ways to test this idea. First, cultures may be systematically related to the national identity of the colonizing power. For example, the British may have implanted a ‘good’ Anglo-Saxon culture into colonies such as Australia and the United States, while the Spanish may have condemned Latin America by endowing it with a Hispanic or Iberian culture (the academic literature is full of ideas like this, for recent versions see Vélez, 1994, North, Summerhill and Weingast, 2000, and Wiarda, 2001). Second, following Landes (1998), Europeans may have had a culture, for example a work ethic or set of beliefs, which was uniquely propitious to prosperity. Finally, following Weber (1930),
Europeans also brought different religions with different implications for prosperity. Such a hypothesis could explain why Latin America is relatively poor since its citizens are primarily Roman Catholic, while North America is relatively rich because its citizens are mostly Protestant.

However, the econometric evidence in Acemoglu, Johnson and Robinson (2001) is not consistent with any these views. Once we control properly for the effects of economic institutions, neither the identity of the colonial power, nor the contemporary fraction of Europeans in the population, nor the proportions of the populations of various religions, are significant determinants of income per capita.

These econometric results are supported by historical examples. For instance, with respect to the identity of the colonizing power, in the 17th century the Dutch had perhaps the best domestic economic institutions in the world but the colonies they created in South-East Asia ended up with institutions designed for the extraction of resources, providing little economic or civil rights to the indigenous population.

It is also be clear that the British in no way simply re-created British institutions in their colonies. For example, by 1619 the North American colony of Virginia had a representative assembly with universal male suffrage, something that did not arrive in Britain itself until 1919. Another telling example is due to Newton (1914) and Kupperman (1993), who showed that the Puritan colony in Providence Island in the Caribbean quickly became just like any other Caribbean slave colony despite the Puritanical inheritance. Although no Spanish colony has been as successful economically as British colonies such as the United States, it is also important to note that Britain had many unsuccessful colonies (in terms of per capita income), such as in Africa, India and Bangladesh (see Acemoglu, Johnson and Robinson, 2004).

To emphasize that the culture or the religion of the colonizer was not at the root of the divergent economic performances of the colonies, Figure 16 shows the reversal among the British colonies (with population density in 1500 on the horizontal axis). Just as in Figure 6, there is a strong negative relationship between population density in 1500 and income per capita today.

With respect to the role of Europeans, Singapore and Hong Kong are now two of the richest countries in the world, despite having negligible numbers of Europeans. Moreover, Argentina and Uruguay have higher proportions of people of European descent than the United States and Canada, but are much less rich. To further document this, Figure 17 shows a similar reversal of fortune for countries where the fraction of those with European descent in 1975 is less than 5 percent of the population.
Overall, the evidence is not consistent with a major role of geography, religion or culture transmitted by the identity of the colonizer or the presence of Europeans. Instead, differences in economic institutions appear to be the robust causal factor underlying the differences in income per capita across countries. Institutions are therefore the fundamental cause of income differences and long-run growth.

5 Why Do Institutions Differ?

We saw that economic institutions matter, indeed are central in determining relative prosperity. In terms of the different fundamental theories that we discussed, there is overwhelming support for the emphasis of North and Thomas on institutions, as opposed to alternative candidate explanations which emphasize geography or culture. Yet, as we discussed in the introduction, finding that differences in economic institutions can account for the preponderance of differences in per-capita income between countries creates as many questions as it answers. For example, why do countries have different economic institutions? If poor countries are poor because they have bad economic institutions why do they not change them to better institutions? In short, to explain the evidence presented in the last two sections we need a theory of economic institutions. The theory will help to explain the equilibrium set of economic institutions in a particular country and the comparative statics of this theory will help to explain why economic institutions differ across countries.

In the Introduction (section 1.2), we began to develop such a theory based on social conflict over economic institutions. We have now substantiated the first point we made there, that economic institutions determine prosperity. We must now move to substantiate our second point, that economic institutions must be treated as endogenous and what which economic institutions emerge depends on the distribution of political power in society. This is a key step towards our theory of economic institutions. In the process of substantiating this point however it is useful to step back and discuss other alternative approaches to developing a theory of economic institutions. Broadly speaking, there are four main approaches to the question of why institutions differ across countries, one of which coincides with the approach we are proposing, the social conflict view. We next discuss each of these separately and our assessment as to whether they provide a satisfactory framework for thinking about differences in economic institutions (see Acemoglu, 2003a, and Robinson, 1998, for related surveys of some of these approaches). We shall conclude that the approach we sketched in section 1.2 is by far the most promising one.
5.1 The Efficient Institutions View—The Political Coase Theorem

According to this view, societies will choose the economic institutions that are socially efficient. How this surplus will be distributed among different groups or agents does not affect the choice of economic institutions. We stress here that the concept of efficiency is stronger than simply Pareto Optimality; it is associated with surplus, wealth or output maximization.

The underlying reasoning of this view comes from the Coase Theorem. Coase (1960) argued that when different economic parties could negotiate costlessly, they will be able to bargain to internalize potential externalities. A farmer, who suffers from the pollution created by a nearby factory, can pay the factory owner to reduce pollution. Similarly, if the current economic institutions benefit a certain group while creating a disproportionate cost for another, these two groups can negotiate to change the institutions. By doing so they will increase the size of the total surplus that they can divide between themselves, and they can then bargain over the distribution of this additional surplus.

Many different versions of the efficient economic institutions view have been proposed. Indeed, assuming that existing economic institutions are efficient is a standard methodological approach of economists, i.e., observing an institution, one tries to understand what are the circumstances that lead it to be efficient. Demsetz (1967) argued that private property emerged from common property when land became sufficiently scarce and valuable that it was efficient to privatize it. More recently, Williamson’s (1985) research, as well as Coase’s (1936) earlier work and the more formal analysis by Grossman and Hart (1986), argues that the governance of firms or markets is such as to guarantee efficiency (given the underlying informational and contractual constraints). Williamson argued that firms emerged as an efficient response to contractual problems that plague markets, particularly the fact that there may be ex-post opportunism when individuals make relationship specific investments. Another famous application of the efficient institutions view is due to North and Thomas (1973) who argued that feudal economic institutions, such as serfdom, were an efficient contract between serfs and lords. The lords provided a public good, protection, in exchange for the labor of the serfs on their lands. In this view, without a modern fiscal system this was an efficient way to organize this exchange. (See Townsend, 1993, for a recent version of the idea that other economic institutions of Medieval Europe, such as the open field system, were efficient).

Williamson and North and Thomas do not specify how different parties will reach agreement to achieve efficient economic institutions, and this may be problematical in
the sense that many economic institutions relevant for development are collective choices not individual bargains. There may therefore be free riding problems inherent in the creation of efficient economic institutions. Nevertheless, the underlying idea, articulated by Becker (1960) and Wittman (1989), is that, at least in democracies, competition among pressure groups and political parties will lead to efficient policies and collective choices. In their view, an inefficient economic institution cannot be stable because a political entrepreneur has an incentive to propose a better economic institution and with the extra surplus generated will be able to make himself more attractive to voters. The efficient institutions view regards the structure of political institutions or power as irrelevant. This may matter for the distribution of total surplus, but it will not matter for efficiency itself. The ‘efficient’ set of political institutions is therefore indeterminate.

The notion that a Coasian logic applies in political life as well as in economics is referred to by Acemoglu (2003a) as the Political Coase Theorem. Although the intuition that individuals and groups will strive towards efficient economic outcomes is appealing, there are both theoretical and empirical limits to the Political Coase Theorem. First, as argued by Acemoglu (2003a) and further discussed below, in politics there is an inherent commitment problem, often making the Political Coase Theorem inapplicable.

Second, the Political Coase Theorem does not take us very far in understanding the effect of economic (or indeed political) institutions on economic outcomes – in this view, economic institutions are chosen efficiently, and all societies have the best possible economic institutions given their needs and underlying structures; hence, with the Political Coase Theorem, economic institutions cannot be the fundamental cause of income differences. However, the empirical results we discussed above suggest a major role for such institutional differences.

The only way to understand these patterns is to think of economic institutions varying for reasons other than the underlying needs of societies. In fact, the instrumental variables and natural experiment strategies we exploited above make use precisely of a source of variation unrelated to the underlying needs of societies. For example, South and North Korea did not adopt very different economic systems because they had different needs, but because different systems were imposed on them for other exogenous reasons. In sum, we need a framework for understanding why certain societies consistently end up with economic institutions that are not, from a social point of view, in their best interests. We need a framework other than the Political Coase Theorem.
5.2 The Ideology View

A second view is that economic institutions differ across countries because of ideological differences—because of the similarity between this and the previous view, Acemoglu (2003a) calls this the Modified Political Coase Theorem. According to this view, societies may choose different economic institutions, with very different implications, because they—or their leaders—disagree about what would be good for the society. According to this approach, there is sufficient uncertainty about the right economic institutions that well-meaning political actors differ about what’s good for their own people. Societies where the leaders or the electorate turn out to be right ex post are those that prosper. The important point is that, just as with the efficient institutions view, there are strong forces preventing the implementation of policies that are known to be bad for the society at large.

Several theoretical models have developed related ideas. For example, Piketty (1995) examined a model where different people have different beliefs about how much effort is rewarded in society. If effort is not rewarded then taxation generates few distortions and agents with such beliefs prefer a high tax rate. On the other hand if one believes that effort is rewarded then low taxes are preferable. Piketty showed that dispersion of beliefs could create dispersion of preferences over tax rates, even if all agents had the same objective. Moreover, incorrect beliefs could be self-fulfilling and persist over time because different beliefs tend to generate information consistent with those beliefs. Romer (2003) also presents a model where voters have different beliefs and showed that if mistakes are correlated, then society can choose a socially inefficient outcome. These models show that if different societies have different beliefs about what is socially efficient they can rationally choose different economic institutions.

Belief differences clearly do play a role in shaping policies and institutions. Several interesting examples of this come from the early experience of independence in former British colonies. For example, it is difficult to explain Julius Nyerere’s policies in Tanzania without some reference to his and other leading politicians’ beliefs about the desirability of a socialist society. It also appears true that in India the Fabian socialist beliefs of Jawaharlal Nehru were important in governing the initial direction that Indian economic policies took.

Nevertheless, the scope of a theory of institutional divergence and comparative development based on ideology seems highly limited. Can we interpret the differences in institutional development across the European colonies or the divergence in the eco-
nomic institutions and policies between the North and South of Korea as resulting from differences in beliefs? For example, could it be the case that while Rhee, Park, and other South Korean leaders believed in the superiority of capitalist institutions and private property rights enforcement, Kim Il Sung and Communist Party members in the North believed that communist policies would be better for the country?

In the case of South versus North Korea, this is certainly a possibility. However, even if differences in beliefs could explain the divergence in economic institutions in the immediate aftermath of separation, by the 1980s it was clear that the communist economic policies in the North were not working. The continued effort of the leadership to cling to these policies, and to power, can only be explained by leaders looking after their own interests at the expense of the population at large. Most likely, North Korean leaders, the Communist Party, and bureaucratic elites are prolonging the current system, which gives them greater economic and political returns than the alternative, even though they fully understand the costs that the system imposes on the North Korean people.

Differences in colonial policies are even harder to explain on the basis of differences in ideology. British colonists established different economic institutions in very different parts of the world: in the Caribbean they set up plantation societies based on slavery, supported by highly oppressive economic institutions. In contrast, the economic institutions that developed in areas where the British settled, and where there was no large population of indigenous to be captured and put to work, and where slavery could not be profitably used, such as northeastern United States, Canada, Australia and New Zealand, were very different. Moreover, differences in the incentives of the colonists in various colonies are easy to understand: when they did not settle, they were choosing economic institutions simply to extract resources from the native population. When they settled in large numbers, economic institutions and policies emerged in order to protect them in the future and encourage investment and prosperity.

These considerations make us tend towards a view which emphasizes the actions of key economic and political agents that are taken rationally and in recognition of their consequences, not simply differences in beliefs. We do not deny that belief differences and ideology often play important roles but we do not believe that a satisfactory theory of institutional differences can be founded on differences in ideology.

5.3 The Incidental Institutions View

The efficient institutions view is explicitly based on economic reasoning: the social costs and benefits of different economic institutions are weighed against each other to deter-
mine which economic institutions should prevail. Efficiency arises because individuals ultimately calculate according to social costs and benefits. Institutions are therefore choices. A different approach, popular among many political scientists and sociologists, but also some economists, is to downplay choices and to think of institutions, both economic and political, as the by-product or unintended consequence of other social interactions or historical accidents. In other words, historical accidents at critical junctures determine institutions, and these institutions persist for a long time, with significant consequences.

Here, we discuss two such theories. The first is the theory of political institutions developed by Moore (1966) in his *Social Origins of Dictatorship and Democracy*, the second is the recent emphasis in the economics literature on legal origins, for example as in the work of Shleifer and his co-authors (La Porta, Lopez-de-Silanes, Shleifer and Vishny, 1998, 1999, Djankov, LaPorta, Lopez-de-Silanes and Shleifer, 2002, 2003, Glaeser and Shleifer, 2002).

Moore attempted to explain the different paths of political development in Britain, Germany and Russia. In particular, he investigated why Britain evolved into a democracy, while Germany succumbed to fascism and Russia had a communist revolution. Moore stressed the extent of commercialization of agriculture and resulting labor relations in the countryside, the strength of the ‘bourgeoisie,’ and the nature of class coalitions. In his theory, democracy emerged when there was a strong, politically assertive, commercial middle class, and when agriculture had commercialized so that there were no feudal labor relations in the countryside. Fascism arose when the middle classes were weak and entered into a political coalition with landowners. Finally, a communist revolution resulted when the middle classes were non-existent, agriculture was not commercialized, and rural labor was repressed through feudal regulations. In Moore’s theory, therefore, class coalitions and the way agriculture is organized determine which political institutions will emerge. However, the organization of agriculture is not chosen with an eye to its effects on political institutions, so these institutions are an unintended consequence. Although Moore was not explicitly concerned with economic development, it is a direct implication of his analysis that societies may end up with institutions that do not maximize income or growth, for example, when they take the path to communist revolution.

Beginning with the work on shareholder rights (La Porta, Lopez-de-Silanes, Shleifer and Vishny, 1998), continuing to the efficiency of government (La Porta, Lopez-de-Silanes, Shleifer and Vishny, 1999) and more recently the efficiency of the legal system
Shleifer and his co-authors have argued that a central source of variation in many critical economic institutions is the origin of the legal system. For example, “Civil laws give investors weaker legal rights than common laws do, independent of the level of per-capita income. Common-law countries give both shareholders and creditors—relatively speaking—the strongest, and French-civil-law countries the weakest, protection.” (La Porta et al., 1998, p. 1116)

These differences have important implications for resource allocation. For example, when shareholders have poor protection of their rights, ownership of shares tends to be more highly concentrated. Djankov et al. (2003) collected a cross-national dataset on how different countries legal systems dealt with the issue of evicting a tenant for nonpayment of rent and collecting on a bounced check. They used these data to construct an index of procedural formalism of dispute resolution for each country and showed that such formalism was systematically greater in civil than in common law countries, and is associated with higher expected duration of judicial proceedings, less consistency, less honesty, less fairness in judicial decisions, and more corruption. Legal origins therefore seems to matter for important institutional outcomes.

Where do legal origins come from? The main argument is that they are historical accidents, mostly related to the incidence of European colonialism. For example, Latin American countries adopted the Napoleonic codes in the nineteenth century because these were more compatible with their Spanish legal heritage. Importantly, the fact that Latin American countries therefore have ‘French legal origin’ is due to a historical accident and can be treated as exogenous with respect to current institutional outcomes.

What about the difference between common law and civil law? Glaeser and Shleifer (2002) argue that the divergence between these systems stems from the medieval period and reflects the balance of power between the lords and the king in England and France. Once these systems established, they persisted long after the initial rationale vanished.

Although we believe that historical accidents and persistence are important, in reality the aspect of choice over institutions seems too important to be denied. Even if institutions have a tendency to persist, their persistence is still a choice, in the sense that if the agents decided to change institutions, change would be possible. There are important examples from history of countries radically changing their legal systems such as in Japan after the Meiji restoration, Russia after the Crimean War, and Turkey under Mustafa Kemal in the 1920’s. Another example might be central planning of the economy. Though many countries adopted this way or organizing the economy some abandoned it while others, such as North Korea and Cuba, still maintain it. The point
here is that though institutions may in some circumstances be the incidental outcome of history, at some point people will start to ask why society has the institutions that it does and to consider other alternatives. At this point we are back in the realm of choice.

5.4 The Social Conflict View

According to this view, economic (and political) institutions are not always chosen by the whole society (and not for the benefit of the whole society), but by the groups that control political power at the time (perhaps as a result of conflict with other groups). These groups will choose the economic institutions that maximize their own rents, and the economic institutions that result may not coincide with those that maximize total surplus, wealth or income. For example, economic institutions that enforce property rights by restricting state predation may not be in the interest of a ruler who wants to appropriate assets in the future. By establishing property rights, this ruler would be reducing his own future rents, so may well prefer economic institutions other than enforced private property. Therefore, equilibrium economic institutions will not be those that maximize the size of the overall pie, but the slice of the pie taken by the powerful groups.

The first systematic development of this point of view in the economics literature is North (1981), who argued in the chapter on “A Neoclassical Theory of the State” that agents who controlled the state should be modeled as self-interested. He then argued that the set of property rights that they would choose for society would be those that maximized their payoff and because of ‘transactions costs,’ these would not necessarily be the set that maximized social welfare. One problem with North’s analysis is that he does not clarify what the transactions costs creating a divergence between the interests of the state and the citizens are. Here, we will argue that commitment problems are at the root of this divergence.

The notion that elites, i.e., the politically powerful, may opt for economic institutions which increase their incomes, often at the expense of society, is of course also present in much of the Marxist and dependency theory literature. For example, Dobb (1948), Brenner (1976, 1982) and Hilton (1981) saw feudalism, contrary to North and Thomas’s (1973) model, as a set of institutions designed to extract rents from the peasants at the expense of social welfare. Dependency theorists such as Williams (1944), Wallerstein (1974-1982), Rodney (1972), Frank (1978) and Cardoso and Faletto (1979) argued that

---

Postan (1966, pp. 603-604) famously estimated that lords extracted about 50% of the entire production of peasants.
the international trading system was designed to extract rents from developing countries to the benefit of developed countries.

The social conflict view includes situations where economic institutions may initially be efficient for a set of circumstances but are no longer efficient once the environment changes. For example, Acemoglu, Aghion and Zilibotti (2001) show that though certain sorts of organizations may be useful for countries a long way from the technological frontier, it may be socially efficient to change them subsequently. This may not happen however because it is not privately rational. An interesting example may be the large business enterprises (the *chaebol*) of South Korea. In the context of political institutions, one might then develop a similar thesis. Certain sets of institutions are efficient for very poor countries but they continue to apply even after they cease to be the efficient institutional arrangement.

In stark contrast to the efficient institutions view, political institutions play a crucial role in the social conflict view. Which economic institutions arise depends on who has political power to create or block different economic institutions. Since political institutions play a central role in the allocation of such power they will be an intimate part of a social conflict theory of economic institutions.

What distinguishes the social conflict view from the ideological view is that social conflict can lead to choices of economic institutions which cause underdevelopment even when all agents have common knowledge that this is so. What distinguishes it from the incidental view is that it emphasizes that institutional choices which cause underdevelopment are conscious choices, rather than the result of some historical accident. The aspect that distinguishes the social conflict view from the efficient institutions view is that it does not assume that institutions are always efficient. This is one possible outcome but it is not the only one or indeed the most likely. Why is this? Why cannot efficiency be separated from distribution? We discuss this issue in the next section.

6 Sources of Inefficiencies

Having motivated our first two assertions in section 1.2, we are now in a position to discuss the third, related to the importance of commitment problems. The inability to commit to how political power will be used in the future means that the impact of economic institutions on efficiency cannot be separated from their effects on distribution.\(^\text{10}\)

In any market situation where economic exchange takes place, and the quid is sepa-

\(^{10}\) An alternative approach would be to stress informational asymmetries (Farrell, 1987).
rated from the pro quo, issues of commitment will arise. That these issues are of crucial importance has been recognized in the literatures on incomplete contracts and renegotiation (e.g., Hart, 1995). Nevertheless, if the legal system functions properly, there is an array of enforceable contracts that owners can sign with managers, workers with employers, borrowers with lenders etc. These contracts can be enforced because there is an authority, a third party, with the power to enforce contracts. Although the authority that is delegated to enforce contracts and to resolve disputes varies depending on the exact situation, all such power ultimately emanates from the state, which, in modern society, has a near-monopoly on the use of legitimate coercion. An owner and manager can write a contract because they believe that the state, and its agents the courts, would be impartial enforcers of the contract.

In contrast, if, for example, a manager believed that the state would be aligned with the interests of the owner and refuse to punish the owner if and when he failed to make a payment stipulated by the contract, then the contract would have little value. Therefore, the presence of an impartial enforcer is important for contracting. The problem when it comes to institutional choices is that there is no such impartial third party that can be trusted to enforce contracts. This is the origin of the commitment problem in politics.

To elaborate on this point, let us consider a situation where society can be governed as a dictatorship or as a democracy. Imagine that the dictator does not relinquish his power, but instead he promises that he will obey the rules of democracy, so that individuals can undertake the same investments as they would in democracy. This promise would not necessarily be credible. As long as the political system remains a dictatorship, there is no higher authority to make the dictator stick to his promise. There is no equivalent of a contract that can be enforced by an impartial third-party. After all, the dictator has the monopoly of military and political power, so he is the final arbiter of conflicting interests. There is no other authority to force the dictator to abide by his promises.

A similar problem plagues the reverse solution, whereby the dictator agrees to a voluntary transition to democracy in return for some transfers in the future to compensate him for the lost income and privileges. Those who will benefit from a transition to democracy would be willing to make such promises, but once the dictator relinquishes his political power, there is no guarantee that citizens would agree to tax themselves in order to make payments to this former dictator. Promises of compensation to a former dictator are typically not credible.

The essence of the problem is commitment. Neither party can commit to compensate the other nor can they commit to take actions that would not be in their interests ex
post. The reason why commitment problems are severe in these examples is because we are dealing with political power. Different institutions are associated with different distributions of political power, and there is no outside impartial party with the will and the power to enforce agreements. In some cases, there may be self-enforcing promises that maintain an agreement. Acemoglu (2003a) discusses such possibilities, but in general, there are limits to such self-enforcing agreements, because they require the participants to be sufficiently patient, and when it comes to matters of political power, the future is uncertain enough that no party would behave in a highly patient manner.

Based on this reasoning, we can now discuss three different channels via which the presence of commitment problems will lead to the choice and persistence of inefficient institutions.

6.1 Holdup

Imagine a situation in which an individual or a group holds unconstrained political power. Also suppose that productive investments can be undertaken by a group of citizens or producers that are distinct from the “political elites”, i.e., the current power holders. The producers will only undertake the productive investments if they expect to receive the benefits from their investments. Therefore, a set of economic institutions protecting their property rights are necessary for investment. Can the society opt for a set of economic institutions ensuring such secure property rights? The answer is often no (even assuming that “society” wants to do so).

The problem is that the political elites—those in control of political power—cannot commit to respect the property rights of the producers once the investment are undertaken. Naturally, ex ante, before investments are undertaken, they would like to promise secure property rights. But the fact that the monopoly of political power in their hands implies that they cannot commit to not hold-up producers once the investments are sunk.

This is an obvious parallel to the hold-up problem in the theory of the firm, where once one of the parties in a relationship has undertaken investments specific to the relationship, other parties can hold her up, and capture some of the returns from her investments. As in the theory of the firm, the prospect of hold-up discourages investment. But now the problem is much more severe, since it is not only investments that are specific to a relationship that are subject to hold-up, but all investments.

This is therefore an example of how inefficient economic institutions arise because of a monopoly of political power. Those with political power cannot commit not to use their
political power ex post, and this translates directly into a set of economic institutions that do not provide secure property rights to groups without political power. The consequence is clear: without such protection, productive investments are not undertaken, and opportunities for economic growth go unexploited.

The reason why these inefficient economic institutions persist (or may be the equilibrium institutions of the society) is related to commitment problems. Parallel to our above example of inducing the dictator to relinquish power, there are two ways to introduce secure property rights. First, in principle, political elites could promise to respect property rights. However, mere promises would not be credible, unless backed up by the political elites relinquishing power, and this would mean relinquishing their rents and privileges. Second, political elites can be bought off by the beneficiaries of a system of more secure property rights. This would typically be achieved by a promise of future payments. For example, after investments are undertaken and output is produced, a share can be given to the political elites. But, as pointed out above, there is another, reverse commitment problem here; the beneficiaries of the new regime cannot commit to making the promised payments to the previous political elites.

Many real world examples illustrate the commitment problems involved in limiting the use of political power. In practice, although buying off dictators and persuading them to leave power is difficult, there have been many attempts to do so, usually by trying to guarantee that they will not be persecuted subsequently. One way of doing this is to give them asylum in another country. Nevertheless, such attempts rarely succeed, most likely again because of commitment problems (the new regime cannot commit to abide by its promises). An illustrative example of this is the attempts by the Reagan administration to persuade Jean-Claude (‘Baby Doc’) Duvalier to relinquish power in Haiti in 1986. In the face of a popular uprising and rising social and economic chaos, the Reagan administration, via the intermediation of the Jamaican Prime Minster Edward Seaga, tried to persuade Duvalier to go into exile. He at first agreed and the White House announced his departure on January 30th, but the next day he changed his mind, unsure that he would really be protected, and stayed in Haiti. One month later he was forced into exile in France by the military.

A more common, and in many ways more interesting strategy to induce dictators to relinquish power is to try to structure political institutions so as to guarantee that they will not be punished. Such institutional changes are sometimes important in transitions to democracy. For example, President Pinochet was willing to abide by the results of the 1989 plebiscite he lost in Chile because as a senator the Constitution protected him
from prosecution. It was only when he left the country that he was vulnerable.

Although Pinochet’s experience illustrates an example of structuring political institutions to achieve commitment, to create durable institutions constraining future use of political power is difficult in practise. These difficulties are well illustrated by the transition from white rule in Rhodesia to majority rule in Zimbabwe. Facing an unwinnable guerilla war, the white elite in Rhodesia sought to negotiate a transition of majority rule, but with enough institutional safeguards that their rents would be protected. These safeguards included the electoral system they wanted, which was used for the first post-independence elections, and massive over-representation in parliament (Reynolds 1999, p. 163). Whites were guaranteed 20% of the seats in the legislature for seven years despite making up only 2-3% of the population and were guaranteed 10 seats of the 40 seat senate. Clauses of the 1980 Constitution were also aimed at directly guaranteeing the property rights of the whites. In particular land reform was outlawed for 10 years after which it could only take place if compensated.

The white negotiators at the Lancaster House talks in 1979 that produced these agreements understood that any promises made by the black majority negotiators about what would happen after independence could not be believed. They sought therefore to find a set of rules that would get around this problem (Herbst, 1990, pp. 13-36). Nevertheless, these guarantees were not enough to protect the property rights (and rents) of the whites in anything other than the short run. The Mugabe regime quickly absorbed the other factions from among the African guerilla opposition, and more moderate relatively pro-white groups, such as Abel Muzorewa’s United African National Council, crumbled. In 1985 the Mugabe regime switched back to the electoral system it preferred (Reynolds, 1999, p. 164) and in 1987, at the first possible opportunity, it removed the guaranteed representation for whites. Though in 1987 Mugabe nominated white candidates for these seats (Horowitz, 1991, pp. 135-136), this did not last for long. In 1990 the senate was abolished. Finally, in 1990 the Constitution was amended to allow for the redistribution of land. Since this time the Mugabe government has begun a sustained policy of land redistribution away from whites through legal and extra-legal means.

6.2 Political losers

Another related source of inefficient economic institutions arises from the desire of political elites to protect their political power. Political power is the source of the incomes, rents, and privileges of the elite. If their political power were eroded, their rents would decline. Consequently, the political elite should evaluate every potential economic change

41
not only according to its economic consequences, such as its effects on economic growth and income distribution, but also according to its political consequences. Any economic change that will erode the elites’ political power is likely to reduce their economic rents in the long run.

As an example, imagine a change in economic institutions that will increase economic growth, but in doing so, will also enrich groups that could potentially contest political power in the future. Everything else equal, greater economic growth is good for those holding political power. It will create greater returns on the assets that they possess, and also greater incomes that they can tax or expropriate. However, if their potential enemies are enriched, this also means greater threats against their power in the future. Fearing these potential threats to their political power, the elites may oppose changes in economic institutions that would stimulate economic growth.

That the threat of becoming a political loser impedes the adoption of better institutions is again due to a commitment problem. If those who gained political power from institutional change could promise to compensate those who lost power then there would be no incentive to block better institutions.

There are many historical examples illustrating how the fear of losing political power has led various groups of political and economic elites to oppose institutional change and also introduction of new technologies. Perhaps the best documented examples come from the attitude of the elites to industrialization during the nineteenth century (see Acemoglu and Robinson, 2000b, 2002). There were large differences between the rates at which countries caught up with British industrialization with many countries completely failing to take advantage of the new technologies and opportunities. In most of these cases, the attitudes of political elites towards industrialization, new technology and institutional change appear to have been the decisive factor, and these attitudes were driven by their fears of becoming political losers. These issues are best illustrated by the experiences of Russia and Austria-Hungary.

In both Russia and Austria-Hungary, absolutist monarchies feared that promoting industrialization would undermine their political power. In Russia, during the reign of Nikolai I between 1825 and 1855 only one railway line was built in Russia, and this was simply to allow the court to travel between Moscow and St. Petersburg. Economic growth and the set of institutions that would have facilitated it were opposed since, as Mosse (1992) puts it “it was understood that industrial development might lead to social and political change.” In a similar vein, Gregory (1991) argues: “Prior to the about face in the 1850’s, the Russian state feared that industrialization and modernization would
concentrate revolution minded workers in cities, railways would give them mobility, and education would create opposition to the monarchy.”

It was only after the defeat in the Crimean War that Nikolai’s successor, Alexsandr II, initiated a large scale project of railway building and an attempt to modernize the economy by introducing a western legal system, decentralizing government, and ending feudalism by freeing the serfs. This period of industrialization witnessed heightened political tensions, consistent with the fears of the elites that times of rapid change would destabilize the political status quo and strengthen their opposition (McDaniel, 1988, gives a detailed account of these events, see also Mosse, 1958).

The consensus view amongst historians also appears to be that the main explanation for the slow growth of Austria-Hungary in the nineteenth century was lack of technology adoption and institutional change, again driven by the opposition of the state to economic change. This view was proposed by Gerschenkron who argued that the state not only failed to promote industrialization, but rather, “economic progress began to be viewed with great suspicion and the railroads came to be regarded, not as welcome carriers of goods and persons, but as carriers of the dreaded revolution. Then the state clearly became an obstacle to the economic development of the country” (1970, p. 89). See also Gross (1973).

The analysis of Fruedenberger (1967, pp. 498-499) is similar. As with the Tsar, the Hapsburg emperors opposed the building of railways and infrastructure and there was no attempt to develop an effective educational system. Blum (1943) pointed to the pre-modern institutional inheritance as the major blockage to industrialization arguing (p. 26) that

“these living forces of the traditional economic system were the greatest barrier to development. Their chief supporter was ... Emperor Francis. He knew that the advances in the techniques of production threatened the life of the old order of which he was so determined a protector. Because of his unique position as final arbiter of all proposals for change he could stem the flood for a time. Thus when plans for the construction of a steam railroad were put before him, he refused to give consent to their execution ‘lest revolution might come into the country’.”
6.3 Economic losers

A distinct but related source of inefficiency stems from the basic supposition of the social conflict view that different economic institutions imply different distributions of incomes. This implies that a move from a bad to a better set of economic institutions will make some people or groups worse off (and will not be Pareto improving). This in turn implies that such groups will have an incentive to block or impede such institutional changes even if they benefit the whole of society in some aggregate sense.

The idea that economic losers impede the choice of efficient economic institutions and economic policies is widespread in economics and was seen earliest in the literature on international trade. Even though free trade may be socially desirable, individuals invested in sectors in which an economy does not enjoy comparative advantage will lose economically from free trade. Since at least the work of Schattshneider (1935) the role of economic losers has been central in understanding why free trade is not adopted. In the context of development economics, this idea was first discussed by Kuznets (1968), developed at length by Olson (1982, 2000) and Mokyr (1990), and formalized by Krusell and Rios-Rull (1996) and Parente and Prescott (1999). Most of the examples discussed in the development literature on economic losers are about technological change—people with specific investments in obsolete technology try to block the introduction of better technology. The most celebrated example is the case of the Luddites, skilled weavers in early nineteenth century England who smashed new mechanized looms which threatened to lead to massive cuts in their wages (see Thomis, 1970, Randall, 1991). Scott (2000, p. 200) relates a similar example from modern Malaysia, “When, in 1976, combine harvesters began to make serious inroads into the wages of poor villagers, the entire region experienced a rash of machine-breaking and sabotage reminiscent of the 1830’s in England.”

That better economic institutions are blocked by individuals whose incomes are threatened by such change is again due to a problem of commitment. If those whose incomes rose when economic institutions changed could promise to compensate those whose incomes fell then there would be no incentive to block better economic institutions. Nevertheless, it is difficult to commit to such transfers. To consider again the example of the Luddites, the factory owners could have promised to pay the weavers high wages in the future even though their skills were redundant. Once the new technology was in place however, owners would have a clear incentive to fire the weavers and hire
much cheaper unskilled workers.\footnote{One possible way round this problem would be for the owners, if they could afford it, to compensate the weavers in advance for their lower future wages. But this would raise the reverse commitment problem: the weavers would have an incentive to take the money and still break the machines – i.e., they could not commit to not blocking the innovations that would reduce their wages even after they had taken the money.}

Although the problem of economic losers is appealing at first sight, has received some attention in the economics literature, and fits into our framework by emphasizing the importance of commitment problems, we view it both theoretically and empirically less important than the holdup and the political loser problems. First, as pointed out in Acemoglu and Robinson (2000b), in theories emphasizing issues of economic losers, there are implicit assumptions about politics, which, when spelled out, imply that political concerns must be important whenever issues of economic losers are present. The idea of economic losers is that certain groups, fearing that they will lose their economic rents, prevent adoption of beneficial economic institutions or technologies. The assumption in this scenario is that these groups have the political power to block socially beneficial changes. But then, if they have the political power to block change, why wouldn’t they allow the change to take place and then use their political power to redistribute some of the gains to themselves? The implicit assumption must therefore be that groups losing economically also experience a reduction in their political power, making it impossible for them to redistribute the gains to themselves after to change takes place. This reasoning therefore suggests that whether certain groups will lose economically or not is not as essential to their attitudes towards change as whether their political power will be eroded. Problems of political losers therefore seem much more important than problems of economic losers.

Possibly for this reason, advocates of the economic losers view have been unable to come up with any well documented examples where the economic losers hypothesis can actually explain first-order patterns of development. For instance, while it is true that the Luddites tried to break machines, they singularly failed to halt the progress of agricultural technology in nineteenth century Britain. The same is true for Malaysia in the 1970s, one of the fastest growing economies in the world at that time. Neither set of workers had sufficient political power to stop change. Indeed, when political powerful groups became economic losers, such as landowners in nineteenth century England who saw land prices and agricultural rents fall rapidly after 1870, they did nothing to block change because their political power allowed them to benefit from efficient economic institutions (Acemoglu and Robinson, 2002).
Perhaps the most interesting failure of economic losers to halt progress in English economic history comes from the impact of the enclosure of common lands. Land has not always been privately owned as property. In much of Africa land is still owned communally, rather than individually, and this was true in Medieval Britain. Starting around 1550 however an ‘enclosure movement’ gathered pace where ‘common land’ was divided between cultivators and privatized. By 1850 this process of enclosures had made practically all of Britain private property.

Enclosure was a heterogenous process (Overton 1996, p. 147) and it also took place at different times in different places. Nevertheless, most of it was in two waves, the so called ‘Tudor enclosures’ between 1550 and 1700 and the ‘parliamentary enclosures’ in the century after 1750.

“From the mid-eighteenth century the most usual way in which common rights were removed was through a specific act of parliament for the enclosure of a particular locality. Such acts ... made the process easier because enclosure could be secured provided the owners of a majority (four-fifths) of the land, the lord of the manor, and the owner of the tithe agreed it should take place. Thus the law of parliament (statue law) only took account of the wishes of those owning land as opposed to the common law which took account of all those who had both ownership rights and use rights to land. Moreover ... in some parishes the ... majority could be held by a single landowner ... parliamentary enclosure often resulted in a minority of owners imposing their will on the majority of farmers.” Overton (1996, p. 158, italics in original)

The historical evidence is unanimous that the incentive to enclose was because “enclosed land was worth more than open common field land ... the general consensus has been that rents doubled” Overton 1996, p. 162). More controversial is the source of this increase in rent. Overton continues (pp. 162-163) “The proportion of profits taken as rents from tenants by landlords is the outcome of a power struggle between the two groups, and the increase in rent with enclosure may simply reflect an increase in landlord power.” Allen (1982, 1992) showed, in his seminal study of the enclosure movement in the South Midlands, that the main impact was a large increase in agricultural rents and a redistribution of income away from those cultivators who had previously used the commons.
The enclosure of common land thus led to a huge increase in inequality in early modern England. Many peasants and rural dwellers had their traditional property rights expropriated. In protest, groups of citizens dispossessed by enclosure attempted to oppose it through collective action and riots—attempting to influence the exercise of political power. These groups were no match for the British state, however. Kett’s rebellion of 1549, the Oxfordshire rebellion of 1596, the Midland Revolt of 1607, and others up to the Swing Riots of 1830-1831 were all defeated (see Charlesworth, 1983). The presence of economic losers did not prevent this huge change in economic institutions and income distribution.

6.4 The Inseparability of Efficiency and Distribution

Commitment problems in the use and the allocation of political power therefore introduce a basic trade-off between efficiency and distribution. For example, when lack of commitment causes hold-ups, those who hold political power know that people will not have the right incentives to invest so growth will be low. In response to this, they might voluntarily give away their power or try to create political institutions that restricted their power. Such a change in political institutions would create better investment incentives. Though this situation is hypothetically possible and has formed the basis for some theories of institutional change (e.g. Barzel, 2001) it appears to be insignificant in reality. Even faced with severe underinvestment, political elites are reluctant to give away their power because of its distributional implications, i.e., because this would reduce their ability to extract rents from the rest of society. Thus poor economic institutions, here lack of property rights and hold-up, persist in equilibrium because to solve the problem, holders of political power have to voluntarily constrain their power or give it away. This may increase the security of property in society and increase incentives to invest, but it also undermines the ability of rulers to extract rents. They may be better off with a large slice of a small pie.

Similar phenomena are at work when there are either political or economic losers. In the first case, namely a situation where political power holders anticipate being political losers, promoting good institutions directly reduces the political power and rents of incumbents and a similar trade-off emerges. Adopting efficient economic institutions will stimulate growth, but when the political status quo is simultaneously eroded the amount of rent accruing to the initially powerful may fall. In the second case, the incomes of those with political power to determine economic institutions falls directly when better economic institutions are introduced. In the absence of credible commitments to side-
payments, those whose incomes fall when better economic institutions are introduced have an incentive to block such institutions.

Because commitment problems seem so endemic in collective choice and politics, it seems natural to believe that institutional change has significant distributional consequences and as a result there will be conflict over the set of institutions in society.

6.5 Comparative Statics

Our analysis so far has made some progress towards our theory of differences in economic institutions. Although our full theory is yet to be developed in the later sections, the different mechanisms discussed in this section already point out the major comparative static implications of our approach regarding when economic institutions protecting the property rights of a broad cross-section of society are likely to be adopted, and when they are likely to be opposed and blocked. We now briefly discuss these comparative statics.

Hold-up, political loser and economic loser considerations lead to some interesting comparative static results which can be derived by considering the political institutions that lie behind these phenomena.

1. First, the perspective of hold-ups immediately suggests that situations in which there are constraints on the use of political power, for example, because there is a balance of political power in society or a form of separation of powers between different power-holders, are more likely to engender an environment protecting the property rights of a broad cross-section of society. When political elites cannot use their political power to expropriate the incomes and assets of others, even groups outside the elite may have relatively secure property rights. Therefore, constraints and checks on the use of political power by the elite are typically conducive to the emergence of better economic institutions.

2. Second, a similar reasoning implies that economic institutions protecting the rights of a broad cross-section are more likely to arise when political power is in the hands of a relatively broad group containing those with access to the most important investment opportunities. When groups holding political power are narrower, they may protect their own property rights, and this might encourage their own investments, but the groups outside the political elites are less likely to receive adequate protection for their investments (see Acemoglu, 2003b).
3. Third, good economic institutions are more likely to arise and persist when there are only limited rents that power holders can extract from the rest of society, since such rents would encourage them to opt for a set of economic institutions that make the expropriation of others possible.

4. Finally, considerations related to issues of political losers suggest that institutional reforms that do not threaten the power of incumbents are more likely to succeed. Therefore, institutional changes that do not strengthen strong opposition groups or destabilize the political situation are more likely to be adopted.

6.6 The Colonial Experience in Light of the Comparative Statics

We now briefly return to the colonial experience, and discuss how the comparative statics discussed here shed light on the differences in economic institutions across the former colonies and the institutional reversal.

The second comparative static result above suggests a reason why better economic institutions developed in places where Europeans settled. In these societies, a relatively broad-based group of Europeans came to dominate political power, and they opted for a set of economic institutions protecting their own property rights. In contrast, in places where Europeans did not settle, especially where they were a small minority relative to a large indigenous population, they did not have the incentives to develop good economic institutions because such institutions would have made it considerably more difficult for them to extract resources from the rest of society.

The third comparative static suggests an important reason why in places with more wealth, resources and also a high density of indigenous population to be exploited, Europeans were more likely to opt for worse institutions, without any protection for the majority of the population, again because such institutions facilitated the extraction of resources by the Europeans.

The first comparative static result, in turn, is related to the persistence of the different types of economic institutions that Europeans established, or maintained, in different colonies. In colonies where Europeans settled in large numbers, they also developed political institutions placing effective checks on economic and political elites. In contrast, the political institutions in colonies with high population density, extractive systems of production, and few Europeans, concentrated power in the hands of the elite, and built a state apparatus designed to use coercion against the majority of the population. These different political institutions naturally implied different constraints on political and
economic elites. In the former set of colonies, there were constraints on the development of economic institutions that would favor a few at the expense of the majority. Such constraints were entirely absent in the latter set of colonies.

Finally, the fourth comparative static is useful in thinking about why many colonies did not attempt to change their economic institutions during the nineteenth century when new economic opportunities made their previous system based on forced labor, slavery, or tribute-taking much less beneficial relative to one encouraging investment in industry and commerce. Part of the answer appears to lie in the fact that the political power of the elites, for example of the plantation owners in the Caribbean, was intimately linked to the existing economic system. A change in the economic system would turn them into political losers, an outcome they very much wanted to avoid.

6.7 Reassessment of the Social Conflict View

So far we have shown that the econometric evidence is convincing that differences in economic institutions are the root cause of differences in prosperity. We then argued that although there are different approaches which can account for variation in economic institutions, the most plausible approach is the social conflict view. Though we believe that there are clear instances where history and ideology matter for the institutional structure of society, and clearly institutions are highly persistent, the most promising approach to understanding why different countries have different institutions is to focus on choices and their subsequent consequences. The social conflict view emphasizes the distributional implication of economic institutions and how commitment problems imply that efficiency and distribution cannot be separated. Hence the fundamental conflict within society over the nature of economic institutions has important implications for economic performance. Some economic institutions will promote growth, but they will not necessarily benefit all groups in society. Alternative economic institutions may induce economic stagnation, but may nevertheless enrich some groups. Which set of institutions results and whether or not a society prospers will be determined by which of these groups has the political power to get the institutions that differentially benefit them. At this point we have therefore substantiated the first three points we made in the introduction. To develop our theory of economic institutions further we need to be more specific about political power—where it comes from and why some people have it and not others. We undertake this task in section 8. Before doing this however the next section discusses three important historical examples of the evolution of economic institutions. We use these examples to show the explanatory power of the social conflict
view and to begin to illustrate in concrete settings how political power works.

7 The Social Conflict View in Action

We now discuss three important examples to bring out the fact that conflict over economic institutions is critical to the functioning of the economy and that this conflict stems, not from differences in beliefs, ideology or historical accidents, but from the impact of economic institutions on distribution. The examples also show that those with political power have a disproportionate effect on economic institutions and they show how the distribution of political power is influenced by different factors. These factors include the allocation of de jure political power through the structure of political institutions and the ability of groups to solve the collective action problem, or exercise what we called de facto political power. With these examples in mind in section 8 we move to discuss in more detail the nature and sources of political power.

7.1 Labor Markets

A market—an opportunity for individuals to exchange a commodity or service—is obviously a fundamental economic institution relevant for development. As Adam Smith (1776) argued, markets allow individuals to take advantage of the benefits of specialization and the division of labor, and scholars such as Pirenne (1937) and Hicks (1969) argued that the expansion of markets was perhaps the driving forces in long-run development.

In the history of Europe a key transformation was from feudal labor market institutions towards modern notions of a free labor market where individuals were able to decide who to work for and where to live. This process of institutional change was intimately connected to the transition from a whole set of feudal economic institutions to the economic institutions we think of as ‘capitalist.’ Most historians see this as key to the economic take-off that began in the nineteenth century. It was the countries which had made the transition away from feudalism most completely, such as England, the Netherlands and France, thanks to the revolution of 1789, which developed most rapidly. It was those where feudalism was still in operation, such as Russia and Austria-Hungary, which lagged far behind.

What can account for this differential evolution of feudalism? Scholars beginning with Postan (1937) saw the demographic collapse caused by the black death in the 1340’s as demolishing feudalism in Western Europe. By dramatically altering the land/labor
ratio as approximately 40% of the population of Europe died (e.g., Cantor, 2001), the Black Death greatly increased the bargaining power of peasants and allowed them to negotiate a free status ending feudal obligations, particularly with respect to labor. Therefore, Postan’s demographic theory implicitly emphasizes the role of political power in the decline of feudalism: this set of economic institutions started to disappear when the political power of the peasants increased and that of lords declined.

In fact, the distribution of power may be even more important in the whole story than Postan’s theory suggests. As first pointed out by Brenner (1976), the demographic theory of the decline feudalism is not consistent with the comparative evidence. Although demographic trends were similar all over Europe and

“it is true that ... in most of Western Europe serfdom was dead by the early sixteenth century. On the other hand, in Eastern Europe, in particular Pomerania, Brandenburg, East Prussia and Poland, decline in population from the late fourteenth century was accompanied by an ultimately successful movement towards imposing extra-economic controls, that is serfdom, over what had been, until then, one of Europe’s freest peasants. By 1500 the same Europe-wide trends had gone a long way towards establishing one of the great divides in European history, the emergence of an almost totally free peasant population in Western Europe, the debasement of the peasantry to unfreedom in Eastern Europe.” (Brenner, 1976, p. 41).

What can explain these divergent outcomes? Brenner notes (p. 51): “It was the logic of the peasant to try to use his apparently improved bargaining position to get his freedom. It was the logic of the landlord to protect his position by reducing the peasants’ freedom.” The outcome “obviously came down to a question of power” (p. 51); whether the peasants or the lords had more political power determined whether serfdom declined or became stronger.

Although we are far from an understanding of the determinants of the relative structure of political power in different parts of Europe, Brenner suggests that an important element was the “patterns of the development of the contending agrarian classes and their relative strength in the different European societies: their relative levels of internal solidarity, their self-consciousness and organization, and their general political resources—especially their relationships to the non-agricultural classes (in particular, potential urban class allies) and to the state” (p. 52). To substantiate this view, Brenner studies how villages tended to be organized differently in Eastern Europe, there was
more of a tendency to individualistic farming; less developed organization of collaborative agricultural practices at the level of the village or between villages; and little of the tradition of the ‘struggle for commons rights’ against the lords which was so characteristic of western development” (p. 57). This differential organization was due to the process of initial occupation of these Eastern lands.

Although many parts of Brenner’s analysis remain controversial, there is general agreement that the decline of feudalism and the transformation of European labor markets were intimately related to the political power of the key groups with opposing interests, the peasants and the lords (see, for example, Aston and Philpin, 1985, on reactions to Brenner’s interpretation). Feudal institutions, by restricting labor mobility and by removing the role of the labor market in allocating labor to jobs, undermined incentives and resulted in underdevelopment. But these same economic institutions created large rents for the aristocracy. As a consequence, aristocracies all over Europe attempted to maintain them. It was when their political power weakened that the process of transformation got underway.

7.2 Financial Markets

Much recent work on growth and development has focused on capital markets. Growth requires investment, so poor agents without access to financial markets will not have the resources to invest. Empirically many scholars have found correlations between the depth of financial markets and growth (see Levine, 2004) and absence of financial markets is at the heart of ambitious theories of comparative development by Banerjee and Newman (1993) and Galor and Zeira (1993).

If the stress on financial markets and financial intermediation is correct, a central issue is to understand why financial systems differ. For example, studies of the development of banking in the United States in the nineteenth century demonstrate a rapid expansion of financial intermediation which most scholars see as a crucial facilitator of the rapid growth and industrialization that the economy experienced. In his recent study Haber (2001, p. 9) found that in the United States, “In 1818 there were 338 banks in operation, with a total capital of $160 million—roughly three times as many banks and bank capital as in 1810. Circa 1860, the United States had 1,579 banks, with a total capital of $422.5 million. Circa 1914 there were 27,864 banks in the United States. Total bank assets totaled $27.3 billion.”

One might see this rapid expansion of banking and financial services as a natural feature. Yet Haber (2001) shows that the situation was very different in Mexico (p. 24).
“Mexico had a series of segmented monopolies that were awarded to a group of insiders. The outcome, circa 1910 could not have been more different: the United States had roughly 25,000 banks and a highly competitive market structure; Mexico had 42 banks, two of which controlled 60 percent of total banking assets, and virtually none of which actually competed with another bank.”

The explanation for this huge difference is not obvious. The relevant technology was certainly readily available everywhere and it is difficult to see why the various types of moral hazards or adverse selection issues connected with financial intermediation should have limited the expansion of banks in Mexico but not the United States. Haber then shows that (p. 9), “at the time that the U.S. Constitution was put into effect in 1789, ... [U.S. banking] was characterized by a series of segmented monopolies that shared rents with state governments via taxes or state ownership of bank stock. In some cases, banks also shared rents directly with the legislators who regulated them.”

This structure, which looked remarkably like that which arose subsequently in Mexico, emerged because state governments had been stripped of revenues by the Constitution. In response, states started banks as a way to generate tax revenues. State governments restricted entry “in order to maximize the amount of rent earned by banks, rent which would then be shared with the state government in the form of dividends, stock distributions, or taxes of various types.”

Thus in the early nineteenth century, U.S. banks evolved as monopolies with regulations aimed at maximizing revenues for the state governments. Yet this system did not last because states began competing among themselves for investment and migrants.

“The pressure to hold population and business in the state was reinforced by a second, related, factor: the broadening of the suffrage. By the 1840s, most states had dropped all property and literacy requirements, and by 1850 virtually all states (with some minor exceptions) had done so. The broadening of the suffrage, however, served to undermine the political coalitions that supported restrictions on the number of bank charters. That is, it created a second source of political competition-competition within states over who would hold office and the policies they would enact.”

The situation was very different in Mexico. After 50 years of endemic political instability the country unified under the highly centralized 40 year dictatorship of Porfirio Diaz until the Revolution in 1910.
In Haber's argument political institutions in the United States allocated political power to people who wanted access to credit and loans. As a result they forced state governments to allow free competitive entry into banking. In Mexico political institutions were very different. There were no competing federal states and the suffrage was highly restrictive. As a result the central government granted monopoly rights to banks who restricted credit to maximize profits. The granting of monopolies turned out to be a rational way for the government to raise revenue and redistribute rents to political supporters (see North, 1981, Chapter 3).

A priori, it is possible that the sort of market regulation Haber found in Mexico might have been socially desirable. Markets never function in a vacuum, but rather within sets of rules and regulations which help them to function. Yet it is hard to believe that this argument applies to Mexico (see also Maurer, 2002). Haber (2001) documents that market regulation was aimed not at solving market failures and it is precisely during this period that the huge economic gap between the United States and Mexico opened up (on which see Coatsworth, 1993, Engerman and Sokoloff, 1997). Indeed, Haber and Maurer (2004) examined in detail how the structure of banking influenced the Mexican textile industry between 1880 and 1913. They showed that only firms with personal contacts with banks were able to get loans. They conclude (p. 5):

“Our analysis demonstrates that textile mills that were related to banks were less profitable and less technically efficient than their competitors. Nevertheless, access to bank credit allowed them to grow faster, become larger, and survive longer than their more productive competitors. The implication for growth is clear: relatively productive firms lost market share to relatively unproductive (but bank-related) competitors.”

Despite the fact that economic efficiency was hurt by regulations, those with the political power were able to sustain these regulations.

7.3 Regulation of Prices

As our final example we turn to the regulation of prices in agricultural markets (which is intimately related to the set of agricultural policies adopted by governments). The seminal study of agricultural price regulation in Africa and Latin America is by Bates (1981, 1989, 1997). Bates (1981) demonstrated that poor agricultural performance in Ghana, Nigeria and Zambia was due to government controlled marketing boards systematically paying farmers prices for their crops much below world levels.
“Most African states possess publicly sanctioned monopsonies for the purchase and export of agricultural goods ... These agencies, bequeathed to the governments of the independent states by their colonial predecessors, purchase cash crops for export at administratively determined domestic prices, and then sell them at the prevailing world market prices. By using their market power to keep the price paid to the farmer below the price set by the world market, they accumulate funds from the agricultural sector” Bates (1981, p. 12).

The marketing boards made surpluses which were given to the government as a form of taxation. Bates (1981, p. 15) notes

“A major test of the intentions of the newly independent governments occurred ... [when] between 1959-1960 and 1961-62, the world price of cocoa fell approximately £50 a ton. If the resources generated by the marketing agencies were to be used to stabilize prices, then surely this was the time to use the funds for that purpose. Instead ... the governments of both Ghana and Nigeria passed on the full burden of the drop in price to the producers.”

Bates continues “Using the price setting power of the monopsonistic marketing agencies, the states have therefore made the producers of cash crops a significant part of their tax base, and have taken resources from them without compensation in the form of interest payments or of goods and services returned.” (pp. 181-9). As a result of this pernicious taxation, reaching up to 70% of the value of the crop in Ghana in the 1970’s, investment in agriculture collapsed as did output of cocoa and other crops. In poor countries with comparative advantage in agriculture such a situation mapped into negative rates of economic growth.

Why were resources extracted in this way? Though part of the motivation was to promote industrialization, the main one is to generate resources that could be either expropriated or redistributed to maintain power

“governments face a dilemma: urban unrest, which they cannot successfully eradicate through co-optation or repression, poses a serious challenge to their interests ... Their response has been to try to appease urban interests not by offering higher money wages but by advocating policies aimed at reducing the cost of living, and in particular the cost of food. Agricultural
policy thus becomes a by-product of political relations between governments and urban constituents” (1981, p. 33)

In contrast to the situation in Ghana, Zambia and Nigeria, Bates (1981, 1989, 1997) showed that agricultural policy in Kenya and Colombia over this period was much more pro-farmer. The difference was due to who controlled the marketing board. In Kenya, farmers were not smallholders, as they were in Ghana, Nigeria and Zambia, and concentrated landownership made it much easier to solve the collective action problem. Moreover, farming was important in the Kikuyu areas, an ethnic group closely related to the ruling political party, KANU, under Jomo Kenyatta (Bates, 1981, p. 122). Farmers in Kenya therefore formed a powerful lobby and were able to guarantee themselves high prices. Even though the government of Kenya engaged in land reform after independence

“80% of the former white highlands were left intact and ... the government took elaborate measures to preserve the integrity of the large-scale farms ... [which] readily combine in defense of their interests. One of the most important collective efforts is the Kenya National Farmer’s Union (KNFU) ... The organization ... is dominated by the large-scale farmers .. [but] it can be argued that the KNFU helps to create a framework of public policies that provides an economic environment favorable to all farmers” Bates (1981, pp. 93-94).

Bates concludes (p. 95) that in Kenya

“large farmers ... have secured public policies that are highly favorable by comparison to those in other nations. Elsewhere the agrarian sector is better blessed by the relative absence of inequality. But is also deprived of the collective benefits which inequality, ironically, can bring.”

In Colombia, farmers were favored because of competition for their votes from the two main political parties. Bates (1997, p. 54) notes

“Being numerous and small, Colombia’s coffee producers, like peasants elsewhere, encountered formidable costs of collective action. In most similar instances such difficulties have rendered smallholders politically powerless. And yet ... Colombia’s peasants elicited favorable policies from politicians,
who at key moments themselves bore the costs of collective action, provisioning the coffee sector with economic institutions and delegating public power to coffee interests.”

How could the coffee growers gain such leverage over national policy?

“A major reason they could do so ... is because the structure of political institutions, and in particular the structure of party competition, rendered them pivotal, giving them the power over the political fortunes of those with ambition for office and enabling them to make or break governments. They thereby gained the power to defeat government officials who sought to orchestrate or constrain their behavior.” Bates (1997, p.51, 54)

A telling piece of evidence in favor of this thesis is that during the 1950’s when a civil war broke out between the two parties, there was five years of military rule and policy turned decisively again the coffee growers, only to switch back again with the peaceful resumption of democracy in 1958.

7.4 Political Power and Economic Institutions

These three examples of the creation of economic institutions have certain features in common. All these institutions, labor market regulation/feudalism, the rules governing financial market development, and agricultural price regulation, clearly reflect the outcome of conscious choices. Feudalism did not end in England for incidental or ideological reasons, but because those who were controlled and impoverished by feudal regulations struggled to abolish them. In Eastern Europe the same struggle took place but with a different outcome. Similarly, Mexico did not end up with different financial institutions than the United States by accident, because of different beliefs about what an efficient banking system looked like, or because of some historical factor independent of the outcome. The same is true for differences in economic policies in Kenya and Ghana. Moreover, different sets of economic institutions arising in different places cannot be argued to be efficient adaptations to different environments. Most historians believe that the persistence of feudal institutions in Eastern Europe well into the nineteenth century explains why it lagged far behind Western Europe in economic development. The difference between the financial institutions of Mexico and the United States also plausibly played a role in explaining why they diverged economically in the nineteenth century. The same holds with respect to agricultural price regulation.
The driving force behind all three examples is that economic institutions are chosen for their distributional consequences. Which specific economic institutions emerge depends on who is able to get their way—who has political power. In England, peasant communities had developed relatively strong local political institutions and were able to consolidate on the shock of the Black Death to put an end to feudal regulations. In Eastern Europe it was the lords who had relatively more power and they were able to intensify feudalism in the face of the same demographic shock (as Domar, 1970, pointed out, the Black Death actually made serfdom more attractive to the lords even if at the same time it increased the bargaining power of the peasants). In the case of banking in the nineteenth century, Haber’s research shows while the authoritarian regime in Mexico had the political power to freely create monopolies and create rents in the banking industry, the United States was different because it was federal and much more democratic. The political institutions of the United States prevented politicians from appropriating the rents that could flow from the creation of monopolies. Finally, in Bates’s analysis, distortionary price regulations arose in Ghana and Zambia, but not in Kenya and Colombia, because in the latter countries agricultural producers had more political power and so could prevent the distortionary policies that would harm their interests.

It is also useful to consider in the context of these examples the mechanisms we discussed in section 6 which underlie the adoption of inefficient economic institutions. Why couldn’t the peasants and lords of feudal Europe negotiate and allow the introduction of a set of economic institutions that would have given peasants incentives to innovate and would have allowed for the efficient allocation of labor? Why couldn’t either the lords have promised not to expropriate any benefits that accrued from innovation, or alternatively the peasants agreed to compensate the lords if feudal labor institutions were abolished? Though it is difficult to find direct evidence on such counterfactuals from the Medieval period, the most plausible explanation is that such ‘deals’ were impossible to make credible. The political power of the lords was intimately connected to feudal institutions and thus dismantling these would not only have increased peasant incentives to innovate, but would also have dramatically altered the balance of political power and the distribution of rents in society. Moreover, under feudal regulations peasants were tied to the land. The introduction of free labor mobility would have given workers an exit option, thus increasing their bargaining power with the lords over the division of output. Thus lords might anticipate being both political and economic losers from the ending of feudalism, even if total output would have increased.
In the case of agricultural price regulation, similar arguments are plausible. Cocoa farmers in Ghana would not have believed promises by governments that they would not expropriate the fruits of higher investment, and the governments themselves would not have believed promises by the farmers to compensate them if they left office. Moreover, efficient sets of economic institutions in Ghana or Nigeria would have strengthened the economic base of the rural sector at the expense of the political power of the then dominant urban sector. Indeed, for Ghana in the 1960’s, we have direct evidence from the urban economy that the threat of being a political loser led to inefficient economic institutions. This emerges in the analysis of Killick (1978, p. 37) of the attempt by the government of Kwame Nkrumah to promote industrialization. Killick notes:

“Even had there been the possibility [of creating an indigenous entrepreneurial class] it is doubtful that Nkrumah would have wanted to create such a class, for reasons of ideology and political power. He was very explicit about this saying ‘we would be hampering our advance to socialism if we were to encourage the growth of Ghanian private capitalism in our midst.’ There is evidence that he also feared the threat that a wealthy class of Ghanaian businessmen might pose to his own political power.”

Further evidence on the importance of political loser considerations comes from E. Ayeh-Kumi one of Nkrumah’s main economic advisers who noted after the coup that Nkrumah (Killick, 1978, p. 60): “informed me that if he permitted African business to grow, it will grow to becoming a rival power to his and the party’s prestige, and he would do everything to stop it, which he actually did.”

In this context, it is interesting that Nkrumah’s solution to consolidate his power was to limit the size of businesses that Ghanaians could own. This caused problems for his industrialization policy which he got round by allowing foreign businessmen to enter Ghana. Though this was inconsistent with his aggressively nationalistic and anti-imperialistic rhetoric, these businessmen did not pose a domestic political threat. Killick (p. 37) notes “Given Nkrumah’s desire to keep Ghanaian private businesses small, his argument that ‘Capital investment must be sought from abroad since there is no bourgeois class amongst us to carry on the necessary investment’ was disingenuous. He goes on to add that, (p. 40) Nkrumah “had no love of foreign capitalists but he preferred to encourage them rather than local entrepreneurs, whom he wished to restrict”.

All these examples show that the distribution of political power in society is crucial for explaining when economic institutions are good and when they are bad. But where
does political power come from and who has political power? In addressing these questions we will develop our theory of economic institutions. In a theory based on social conflict where economic institutions are endogenous, it will be to differences in political institutions and the distribution of political power that we must look to explain variation in economic institutions.

8 Theory of Institutions

8.1 Sources of Political Power

Who has political power and where does it come from? As we noted in the Introduction (section 1.2, point 4), political power comes from two sources. First, an individual or group can be allocated *de jure* power by political institutions. But institutions are not the only source of power. A second type of political power accrues to individuals or groups if they can solve the collective action problem, create riots, revolts, or demonstrations, own guns, etc.. We call this type of power *de facto* political power (see Acemoglu and Robinson, 2003, chapter 5).

Actual political power is the composition, the joint outcome, of *de jure* and *de facto* power. To see how this works out in practice, consider the situation in Chile in the early 1970’s. Salvador Allende was elected President with a majority of the popular vote. The formal political institutions of democracy in Chile allocated power to him to propose legislation, issue decrees, etc. Consequently, even though he did not have an absolute majority in congress, Allende had a great deal of *de jure* political power. Political power is not just *de jure* however; it does not simply stem from political institutions. Allende, despite being empowered under the Chilean Constitution, was overthrown by a military coup in 1973. Here, the military under the leadership of General Pinochet, were able to use brute force and guns to over-ride the formal political institutions. The ability to use force is one example of *de facto* political power.

As we suggested in the introduction, the relationship between political power and economic and political institutions is complex and dynamic. Consider the example we discussed in section 7.2, the research by Haber on the comparative financial evolution of Mexico and the United States in the nineteenth century. Haber traced the different evolution of economic institutions to differences in initial political institutions. These political institutions led to different distributions of power and this was critical for the emergence of good financial institutions in the United States, whereby those who benefitted from a competitive banking industry were able to force politicians to provide the
rules which would guarantee it. But where did these differences in political institutions come from? These differences were partly a result of political events in the nineteenth century, and partially a result of different colonial political institutions. In the United States, during the initial phase of colonization in the early seventeenth century. Very low population density and lack of easily exploitable resources forced colonizing companies and the British state to make both economic and political concessions; they granted the settlers access to land and accepted the formation of representative democratic institutions (Morgan, 1975). Consequently, even at independence the United States had relatively democratic political institutions (Keyssar, 2000). Moreover, the initial egalitarian distribution of assets and the high degree of social mobility made for a situation where, at least in the northern states, the distribution of economic resources, and thus de facto power, was relatively equal. The relatively representative political institutions therefore persisted and were supported by the balance of de facto power in society.

In Mexico there were very different initial conditions during the colonial period with a large indigenous population and rich silver mines to exploit. This led to a much more hierarchical and authoritarian balance of political power and very different colonial economic institutions (see Engerman and Sokoloff, 1997, Acemoglu, Johnson and Robinson, 2004). These conditions fed into the different institutional structures at independence, the United States with its constitution, checks and balances and federalism, Mexico with its much more centralized, unchecked, unbalanced and absolutist state. These different political institutions then led to very different economic institutions and economic outcomes after independence. Thus, in some ultimate sense, the source of different political institutions were different initial conditions during the colonial period.

Consider now the evidence presented by Bates. Agricultural policies were better in Kenya because large farmers could solve the collective action problem and exercise de facto political power. But the main reason for the existence of large farms was that British settlers expropriated the land from Africans during the expansion of colonialism (see Berman and Lonsdale, 1992). Thus previous combinations of formal political institutions (colonial institutions) and de facto power (the military might of the British Empire) determined economic institutions, feeding into the future distribution of de facto power even after the nature of de jure power changed dramatically with independence.

We can now see that these examples substantiate the dynamic model that we sketched in section 1.2. There we showed that at any date, political power is shaped by political institutions, which determine de jure power, and the inherited distribution of resources, which affect the balance of de facto power. Political power then determines economic
institutions and economic performance. It also influences the future evolution of political power and prosperity. Economic institutions determine the distribution of resources at that point, which, in turn, influences the distribution of de facto power in the future. Similarly, the distribution of power at any point determines not just the economic institutions then, but also the future political institutions. Thus the allocation of political power at one date, because of the way it influences the distribution of resources and future political institutions, has a crucial effect on the future allocation of both de facto and de jure political power.

Both the comparison Haber made between Mexico and the United States, and that which Bates made between Ghana, Zambia, Kenya and Colombia illustrate this diagram in action. They show how political institutions and de facto power combine to generate different set of economic institutions, how these institutions determine both the distribution of resources and the growth rate of the economy, and how power and institutions evolve over time, often in ways that tend to reinforce particular initial conditions.

8.2 Political Power and Political Institutions

The examples we discussed above showed how political power depends on political institutions and de facto power, and how this determines economic institutions. Moreover, we saw that at any time the pre-existing economic institutions will be an important determinant of the distribution of de facto power. The final element to emphasize is how political institutions evolve over time and how they influence the distribution of political power.

To see why political institutions are so important as a source of political power think of a situation where a group, say the Chilean army in the early 1970's, has a great deal of de facto power. Indeed, it has so much de facto power that it can overrule the Chilean Constitution, making the political institutions largely irrelevant. In fact in Chile the de facto power of the military was able to overthrow the legitimate government and completely reverse the economic policies and economic institutions chosen by the Allende government (including land reform and mass nationalization of industry). Not only did the military reverse the economic institutions preferred by Allende and the groups who elected him, they then implemented their own preferred set of economic institutions, in particularly deregulating the trade regime and the economy. Yet the Pinochet regime was heavily concerned with formal political institutions, and in 1980 Pinochet re-wrote the constitution.

If de facto power was decisive in Chile what is the role for political institutions? If
the constitution can be overthrown, why bother to re-write it? The secret to this lies in the intrinsically transitory nature of de facto power. Yes, the military were able to organize a coup in 1973 but this was only because times were uniquely propitious. There was a world-wide economic crisis, and factions of the military that opposed the coup could be marginalized. Moreover, the United States government at the time was happy to encourage and endorse the overthrow of a socialist government, even if it had been democratically elected. The coming together of such circumstances could not be expected to happen continually, hence once Chilean society re-democratized, as it did after 1990, the military would not be able to continually threaten a coup. In response to this Pinochet changed the political institutions in order to attempt to lock in the power of the military, and thus the economic institutions that he/they preferred. Therefore, the important role for political institutions is that they influence the future allocation of political power. This dynamic role is crucial because it explains the major desire of agents to change political institutions when they get the chance—this is how they can attempt to enduringly alter the balance of political power in their favor (see Acemoglu and Robinson, 2003).

8.3 A Theory of Political Institutions

We now have in place the outlines of our theory of institutions. There are seven points to emphasize, paralleling the discussion in section 1.2 and our diagrammatic exposition there. First, individuals have preferences over economic institutions because of the allocation of resources that these institutions induce.

Second, peoples’ preferences typically do not agree because efficiency and distribution cannot be separated. Different economic institutions will benefit different groups, and this will determine the preferences of these individuals and groups with respect to economic institutions.

Third, the problem of commitment explains why efficiency and distribution are inseparable. Economic institutions are collective choices, and they are chosen and sustained

---

12 The empirical literature on the collective action problem has recognized that the difficulty of solving the collective action problems lead collective action to typically be transitory. Lichbach (1995, p. 17) notes “collective action, if undertaken on a short-term basis, may indeed occur; collective action that requires long periods to time does not ... Given that most people’s commitments to particular causes face inevitable decline, most dissident groups are ephemeral, most dissident campaigns brief.” This transitory nature of collective action is echoed by Tarrow (1991, p. 15) who notes “the exhaustion of mass political involvement,” while Ross and Gurr (1989, p. 414) discuss political “burnout.” Similarly, Hardin (1995, p. 18) argues that “the extensive political participation of civil society receives enthusiastic expression only in moments of state collapse or great crisis. It cannot be maintained at a perpetually high level.”
by the state. Since there is no third party to enforce the decisions of the state, problems of commitment are particularly severe in the political realm.

Fourth, the equilibrium structure of economic institutions will therefore be determined by who has the power to get their way, i.e., who can create and sustain economic institutions that benefit themselves. The distribution of political power thus determines economic institutions, the allocation of resources and the rate of economic growth.

Fifth, political power has two forms: de jure power determined by the political institutions, such as the constitution and the electoral rules, and de facto power, which stems from the ability to solve the collective action problem, mobilize weapons etc.. De facto power can influence political outcomes independently of the political institutions, and its distribution often critically determines how a given set of institutions works in practice and whether or not they are actually obeyed.

Sixth, the distribution of de facto political power at any date is influenced to a large degree by the distribution of resources in society, since those with greater resources can command more power both through legitimate and intimate means, and perhaps can also solve the collective action problem more efficiently. Naturally, the distribution of resources at this point is influenced by economic institutions and economic outcomes in the past.

Finally, political institutions are also endogenous; the current balance of political power, incorporating both de jure and de facto elements, also determines future political institutions. Political institutions are important because they allocate, at least within the limits defined by the exercise of future de facto power, the allocation of future de jure political power. Since de facto power, because of the nature of the collective action problem, is intrinsically transitory and difficult to wield, political institutions are often crucial in creating a source of durable political power. This makes it very attractive for groups to use their de facto political power to change political institutions so as to modify the distribution of future political power in their favor.

9 The Theory in Action

We now consider two examples that demonstrate our theory of institutions in action. Like the examples discussed in section 7, these examples contain all the elements of our theory laid out in a skeletal way in section 1.2. They show the role of political power in determining economic institutions, they demonstrate the different factors, both de facto and de jure, that determine political power, and they illustrate how de facto political power is often used to change political institutions in order to influenced the future
distribution of de jure political power.

9.1 Rise of Constitutional Monarchy and Economic Growth in Early Modern Europe

Our first example is the rise of constitutional monarchy in Europe. In the medieval period most European nations were governed by hereditary monarchies. However, as the feudal world changed, various groups struggled to gain political rights and reduce the autocratic powers of monarchies. In England, this process began as early as 1215 when King John was forced by his barons to sign the Magna Carta, a document which increased the powers of the barons, introduced the concept of equality before the law, and forced subsequent kings to consult with them. Many other European nations also developed ‘parliaments’ which kings could summon to discuss taxation or warfare (see Graves, 2001, Ertman, 1997). Nevertheless, the movement towards limited, constitutional monarchy was not linear or simple. Indeed, in France, certainly from the beginning of Louis XIV’s reign in 1638, a more powerful absolutist monarchy appeared with very few controls. Indeed the feudal French parliaments, the Estates General, were not summoned between 1614 and 1788, just before the Revolution.

In England, the Tudor monarchs, particularly Henry VIII and then Elizabeth I, followed by the first Stuart kings, James I and Charles I, also attempted to build an absolutist monarchy. They failed, however, mostly because of Parliament, which blocked attempts to concentrate power. The constitutional outcome in England was settled by the Civil War from 1642-1651 and the Glorious Revolution in 1688. In the first of these conflicts the forces of Parliament defeated those loyal to Charles I and the king was beheaded. In 1660 the monarchy was restored when Charles II became king, but his brother James II was deposed in 1688 and Parliament invited William of Orange to become king.

Other places in Europe, particularly the Netherlands, saw similar developments to those in England. Under the Dukes of Burgundy, the Netherlands had won a considerable amount of political and economic freedom, particularly under the Grand Privilege of 1477 which gave the States General of the Burgundian Netherlands the right to gather on their own initiative and curbed the right of the ruler to raise taxes. However, the Netherlands were inherited by the Hapsburgs through marriage, and by 1493 Maximilian of Hapsburg had reversed the Grand Privilege. After 1552, war with France increased the Hapsburgs’ fiscal needs and led them to impose a large tax burden on the Netherlands, already a prosperous agricultural and mercantile area. Growing fiscal and religious
resentment in 1572 led to a series of uprisings against the Hapsburgs, mostly orchestrated by commercial interests. These culminated in the War of Independence which was finally won in 1648.

While England and the Netherlands were developing limited constitutional governments, Spain and Portugal were moving in the same direction as France, towards greater absolutism. Davis (1973a, p. 66) notes [in Castille] “the king ruled subject only to weak constitutional restraints. In the first decades of the sixteenth century the crown had reduced the pretensions of the Castillian nobility and towns, so that the representative body, the Cortes, could obstruct but not in the last resort prevent royal tax raising.”

These differential institutional trajectories were of enormous consequence. Netherlands and England moved ahead economically of the rest of Europe precisely because they developed limited, constitutional government. This form of government led to secure property rights, a favorable investment climate and had rapid multiplier effects on other economic institutions, particularly financial markets (see, e.g., North and Weingast, 1989, de Vries and van der Woude, 1997). While the Netherlands and Britain prospered, France was convulsed by the French Revolution, and by the nineteenth century Spain and Portugal were impoverished backward nations. How can we account for these diverging paths in the early modern period? Why did England and the Netherlands develop limited constitutional rule, while France, Spain and Portugal did not?

We proposed an explanation in Acemoglu, Johnson and Robinson (2002b) related to the differential responses of these countries to the opportunities of ‘Atlantic trade’, that is, overseas trade and colonial activity unleashed by the discovery of the New World and the rounding of the Cape of Good Hope at the end of the fifteenth century. All five nations engaged in Atlantic trade, but they did so in different ways, with very different implications for the organization of society, political institutions and subsequent economic growth.

In England “most trade was carried on by individuals and small partnerships, and not by the Company of Merchant Adventurers, the Levant Company ... or others of their kind” (Davis, 1973b, p. 41). At least by 1600 there was quite free entry into the English merchant class. The same was true in the Netherlands. In contrast, Cameron (1993, p. 127) describes the Portuguese situation as follows: “The spice trade in the East Indies of the Portuguese Empire was a crown monopoly; the Portuguese navy doubled as a merchant fleet, and all spices had to be sold through the Casa da India (India House) in Lisbon ... no commerce existed between Portugal and the East except that organized and controlled by the state.” In Spain, similarly, colonial trade was a monopoly of the Crown
of Castille, which they delegated to the Casa de Contratación (House of Trade) in Seville. This merchants guild was closely monitored by the government (Parry, 1966, Ch 2). The main aim of these regulations was to make sure that all of the gold and silver from the Americas flowed back to Spain, creating a source of direct tax revenues for the crown. As a result, Latin American colonies were forbidden to buy manufactured goods from anywhere other than Spain, and all exports and imports had to pass through controlled channels. For example, until the Bourbon reforms of the mid eighteenth century, nothing could be exported directly from Buenos Aires, and if somebody produced anything for export on the Pampas, it had to be carried over the Andes and exported from Lima in Peru!

The source of the differences in the organization of trade, in turn, reflected the different political institutions of these countries. At the time, the granting of trade monopolies was a key fiscal instrument to raise revenues; the more powerful monarchs could increase their revenues by granting trade monopolies or by directly controlling overseas trade, while weaker monarchs could not. At the turn of the fifteenth century, the crown was much stronger in France, Spain and Portugal than in Britain and the Netherlands, and this was the most important factor in the differences in the organization of overseas trade. In fact, when both Tudor and Stuart monarchs attempted to create monopolies similar to those in Spain and Portugal, this was successfully blocked by the English Parliament (see, for example, Hill, 1969). Consequently, as world trade expanded in the sixteenth and early seventeenth centuries, it enriched merchants engaged in overseas trade in England and the Netherlands, but the crown and groups allied with it in France, Spain and Portugal. In England and the Netherlands, but not in France, Spain and Portugal, a new class of merchants (and gentry in England) arose with interests directly opposed to those of the Stuarts and the Hapsburgs, and this group was to play a central part in subsequent political changes.

In the case of the Netherlands, de Vries and van der Woude (1997) argue that “urban economic interests ultimately believed it advantageous to escape the Hapsburg imperial framework” (p. 369), and that it was “the traditional pillars of the maritime economy ... that supported and strengthened the young Republic in its hour of need” (p. 366). Moreover, in the case of Amsterdam, “[Hapsburgs’] opponents included most of the city’s international merchants ... [I]n 1578 a new Amsterdam city council threw the city’s lot in with the Prince of Orange ... among the merchants returning from ... exile were [those whose families] and several generations of their descendents would long dominate the city” (1997, p. 365). The expansion of world trade enriched and
expanded precisely those groups within Dutch society most opposed to Hapsburg rule. Israel (1995, pp. 241-242) writes: “From 1590, there was a dramatic improvement in the Republic’s economic circumstances. Commerce and shipping expanded enormously, as did the towns. As a result, the financial power of the states rapidly grew, and it was possible to improve the army vastly, both qualitatively, and quantitatively, within a short space of time. The army increased from 20,000 men in 1588 to 32,000 by 1595, and its artillery, methods of transportation, and training were transformed” (see also Israel, 1989, Chapter 3). By 1629, the Dutch were able to field an army of 77,000 men, 50% larger than the Spanish army of Flanders (Israel, 1995, p. 507). As a consequence of the Dutch revolt, the Netherlands developed a republican form of government closely attuned to mercantile interests. De Vries and van der Woude (1997, p. 587) describe the new political elite following the Dutch Revolt as: “6 to 8% of urban households with incomes in excess of 1,000 guilders per year. This was the grote burgerij from whom was drawn the political and commercial leadership of the country. Here we find, first and foremost, the merchants,” and point out how merchants dominated the governments of Leiden, Rotterdam and the cities in two largest states, Zeeland and Holland.

In England, the Civil War and Glorious Revolution coincided with the great expansion of English mercantile groups into the Atlantic. The East India Company was founded in 1600 as the culmination of a series of efforts to develop trade routes with Asia. The 1620s saw the great expansion of tobacco cultivation in Virginia. This was shortly followed by the development of the highly profitable English sugar colonies in the Caribbean. Finally, in the 1650s the English began to take over the Atlantic slave trade. Both the Civil War and the Glorious Revolution were at root battles over the rights and prerogatives of the monarchy. In both cases new merchant interests predominantly sided with those in the gentry demanding restrictions on the powers of the monarchy in order to protect their property and commerce.

The majority of merchants trading with the Americas and in Asia supported the Parliament during the Civil War. Brunton and Pennington (1954, p. 62) also note “in the country as a whole there was probably a preponderance of Parliamentarian feeling among merchants.” Detailed analyses of the initial members of the Long Parliament in 1640 show that a significant majority of merchants supported the Parliamentarian cause (see Brenner, 1973, 1993, Keeler, 1954, and Brunton and Pennington, 1954). Members of the Commons from the City of London (the main center of mercantile activity), as well as many non-London commercial constituencies, such as Southampton, Newcastle and Liverpool, supported the Parliament against the King. These men included both
professional merchants and aristocrats who invested in colonizing the Americas. These
new merchants also provided the financial support needed by the Parliament in the
difficult early days of the war. They became the customs farmers for the new regime
and therefore advanced tens of thousands of pounds that were essential in building up
the army (Brenner, 1973, p. 82).

Pincus (1998, 2001, 2002) further documents the critical role of mercantile interests in
actively supported William’s plan for invasion, and provided a key financial prop to
the regime in the critical early months.” He notes that James II favored the East India
Company and granted various monopoly privileges, alienating the merchant class. Thus,
“no wonder the merchant community poured money into William of Orange’s coffers in
1688.” (Pincus, 2002, pp. 32-33).

The changes in the distribution of political power, political institutions and thus
economic institutions that took place in England and the Netherlands had no counter-
parts in countries with relatively absolutist
institutions, like Spain and Portugal, where
the crown was able to closely control the expansion of trade. In these countries it was
the monarchy and groups allied with it that were the main beneficiaries of the early
profits from Atlantic trade, and groups favoring political and economic change did not
become strong enough to induce such change. As a result, only in the Netherlands and
England did constitutional rule emerge, and only in these two countries were property
rights secure. As a result it was these same two countries that prospered.

Why could the monarchies of Spain and Portugal not negotiate a more efficient set of
institutions? Alternatively why did the Stuart monarchs in England have to be beheaded
or forced from power before better economic institutions could emerge?

It seems quite clear that a change to a more efficient set of institutions in Spain and
Portugal would not have been possible under the auspices of the absolutist state, and a
reduction in the power of the state was certainly inimical to the interest of the crown.
In the case of England, Hill (1961a) argues directly that the reason that the Tudor
and Stuart monarchs were not in favor of efficient economic institutions is because they
feared that this would undermine their political power. He notes:

“in general the official attitude to industrial advance was hostile, or at
best indifferent. It was suspicious of social change and social mobility, the
rapid enrichment of capitalists, afraid of the fluctuations of the market and of
unemployment, of vagabondage, and social unrest ... the Elizabethan codes
aimed at stabilizing the existing class structure, the location of industry and
the flow of labor supply by granting privileges and by putting hindrances in the way of the mobility and the freedom of contract.”

The account so far explains why a change in the balance of (de facto) political power in England and the Netherlands led to a set of economic institutions favoring the interests of merchants. But in fact much more happened during the seventeenth century; an entirely new set of political institutions, constitutional regimes, restricting the power of the monarchy, were introduced. The reason why the merchants and the gentry in England (and the merchants in the Netherlands) used their newfound powers for political reform illustrates the dynamics of political power emphasized by our theoretical framework.

For example in the case of England, although in 1688 the Parliament might have been strong, it could not be sure that this power would endure. Indeed, the ability to solve the collective action problem and wield de facto power is intrinsically transitory. For instance, the Parliament vanquished James II with the help of a Dutch army, after which they invited William of Orange to take the throne. But how could they anticipate whether or not William would try to assert the absolutist prerogatives that James II had demanded?

The way to make transitory power permanent is to embody it into the rules of the game which is exactly what the English Parliament did after 1688. The changes in institutions after 1688 had large and important effects. For instance, in the eighteenth century the English monarchy was able to borrow huge amounts of money because the fiscal control of Parliament guaranteed that it would not default (see Brewer, 1988, Stasavage, 2003). This borrowing has been seen as crucial to the success of the English war machine. Moreover, with the Parliament in control of fiscal policy, the crown was never able to raise money through arbitrary taxation and not able any more to grant monopoly rights in exchange for money—an issue which had previously been a constant source of friction between the English crown and Parliament. Similarly, after 1688, the greater security of property rights in England led to a huge expansion of financial institutions and markets (Neal, 1990), which, North and Weingast (1989) argue, laid the institutional foundations for the Industrial Revolution.

Of course the English crown was not without some residual power and might have attempted to mount a coup against the Parliament to change political institutions back in its favor. This certainly happened in some places, such as in France after 1849 when Louis Napoleon mounted a successful coup to restore absolutist privileges lost in 1848. Nevertheless, changes in political institutions altered the nature of the status quo in
significant ways, and therefore, influenced the future distribution of de jure political power. Political institutions are not cast stone, and they can change, but they still create a source of political power more durable than mere de facto power.

9.2 Summary

The emergence of constitutional rule in some societies of early modern Europe therefore provides a nice example of how economic institutions, which shape economic outcomes, are determined by political power, which is in turn determined by political institutions and the distribution of resources in society. The Netherlands and England prospered in this period because they had good economic institutions, particularly secure property rights and well developed financial markets. They had these economic institutions because their governments were controlled by groups with a strong vested interest in such economic institutions. These groups wielded political power because of the structure of political institutions, i.e., they received de jure power in the Netherlands after the Dutch Revolt and in England after the Civil War and Glorious Revolution.

Moving one step back, we see that political institutions allocated more de jure political power to commercial interests in England and the Netherlands than in France, Spain and Portugal because of major changes in political institutions during the 1600s. These changes took place because commercial interests in England and the Netherlands acquired significant de facto political power as a result of their improving economic fortunes, itself a consequence of the interaction of Atlantic trade and the organization of overseas trade in these countries. Crucially for our framework, these commercial interests used their de facto power to reform (or revolutionize) political institutions so as to acquire de jure political power and solidify their gains.

These events, therefore, illustrate the various elements of our theoretical framework. In particular, they show how it is useful to think of political institutions and the distribution of economic resources as the state variables of the dynamic system, which determine the distribution of political power, and via this channel, economic institutions and economic outcomes. Political institutions and the distribution of economic resources are, themselves, endogenous, determined by political power and economic institutions, as exemplified by the fact that the distribution of economic resources changed significantly during the sixteenth century as a result of the new economic opportunities presented by the rise of Atlantic trade, and these changes were crucially influenced by the existing economic institutions (the organization of overseas trade). Furthermore, the change in the balance of political power led to the changes in political institutions through the
English Civil War, the Glorious Revolution and the Dutch Revolt.

9.3 Rise of Electoral Democracy in Britain

Our second example, based on Acemoglu and Robinson (2000a, 2001, 2003), is the rise of mass democracy. In the early nineteenth century, European countries were run by small elites. Most had elected legislatures, often descendents of medieval parliaments, but the franchise was highly restricted to males with relatively large amounts of assets, incomes or wealth. However, as the century and the Industrial Revolution progressed, this political monopoly was challenged by the disenfranchised who engaged in collective action to force political change.

In response to these developments, the elites responded in three ways. First by using the military to repress the opposition, as in the responses to the revolutions of 1848. Second, by making concessions to buy off opposition—this is the standard explanation for the beginnings of the welfare state in Germany under Bismarck. Finally, if neither repression nor concessions were attractive or effective, elites expanded the franchise and gave political power to the previously disenfranchised—they created the precedents of modern democracy.

The history of the rise of democracy in Britain is in many ways representative of the experiences of many other European countries. The first important move towards democracy in Britain came with the First Reform Act of 1832. This act removed many of the worst inequities under the old electoral system, in particular the ‘rotten boroughs’ where several members of parliament were elected by very few voters. The 1832 reform also established the right to vote based uniformly on the basis of property and income. The reform was passed in the context of rising popular discontent at the existing political status quo in Britain.

By the 1820s the Industrial Revolution was well under way and the decade prior to 1832 saw continual rioting and popular unrest. Notable were the Luddite Riots from 1811-1816, the Spa Fields Riots of 1816, the Peterloo Massacre in 1819, and the Swing Riots of 1830 (see Stevenson, 1979, for an overview). Another catalyst for the reforms was the July revolution of 1830 in Paris. Much of this was led and orchestrated by the new middle-class groups who were being created by the spread of industry and the rapid expansion of the British economy. For example, under the pre-1832 system neither Manchester nor Sheffield had any members of the House of Commons.

There is little dissent amongst historians that the motive for the 1832 Reform was to avoid social disturbances (e.g., Lang, 1999, p. 36). The 1832 Reform Act increased
the total electorate from 492,700 to 806,000, which represented about 14.5% of the adult male population. Yet, the majority of British people (the remaining 23 million) could not vote, and the elite still had considerable scope for patronage, since 123 constituencies still contained less than 1,000 voters. There is also evidence of continued corruption and intimidation of voters until the Ballot Act of 1872 and the Corrupt and Illegal Practices Act of 1883. The Reform Act therefore did not create mass democracy, but rather was designed as a strategic concession. In presenting his electoral reform to the British Parliament in 1831, the Prime Minister Earl Grey was well aware that this was a measure necessary to prevent a likely revolution. He argued:

“The Principal of my reform is to prevent the necessity for revolution ... reforming to preserve and not to overthrow.” (quoted in Evans, 1983, p. 212).

Unsurprisingly therefore, the issue of parliamentary reform was still very much alive after 1832, and it was taken up centrally by the Chartist movement. But as Lee (1994, p. 137) notes “The House of Commons was largely hostile to reform because, at this stage, it saw no need for it.” This had changed by 1867, largely due to a juxtaposition of factors, including the sharp business cycle downturn that caused significant economic hardship and the increased threat of violence. Also significant was the founding of the National Reform Union in 1864 and the Reform League in 1865, and the Hyde Park riots of July 1866 provided the most immediate catalyst.

Lang (1999, p. 75) sums up his discussion by saying “The Hyde Park affair, coupled with other violent outbursts, helped to underscore the idea that it would be better to keep the goodwill of the respectable workers than to alienate them.” Reform was initially proposed by the Liberal Prime Minister Russell in 1866 but was defeated by the Conservatives and dissident MP’s. As a result Russell’s government fell, and the Conservatives formed a minority administration with Lord Derby as their leader in the House of Lords, and Disraeli in charge of the House of Commons. It was Disraeli who then constructed a coalition to pass the Second Reform Act in 1867. As a result of these reforms, the total electorate was expanded from 1.36 million to 2.48 million, and working class voters became the majority in all urban constituencies. The electorate was doubled again by the Third Reform Act of 1884, which extended the same voting regulations that already existed in the boroughs (urban constituencies) to the counties (electoral constituencies in the rural areas). The Redistribution Act of 1885 removed many remaining inequalities in the distribution of seats and from this point on Britain
only had single member electoral constituencies (previously many constituencies had elected two members—the two candidates who gained the most votes). After 1884 about 60% of adult males were enfranchised. Once again social disorder appears to have been an important factor behind the 1884 act.

In Britain, the Reform Acts of 1867-1884 were a turning point in the history of the British state. Economic institutions also began to change. In 1871 Gladstone reformed the civil service, opening it to public examination, making it meritocratic. Liberal and Conservative governments introduced a considerable amount of labor market legislation, fundamentally changing the nature of industrial relations in favor of workers. During 1906-1914, the Liberal Party, under the leadership of Asquith and Lloyd George, introduced the modern redistributive state into Britain, including health and unemployment insurance, government financed pensions, minimum wages, and a commitment to redistributive taxation. As a result of the fiscal changes, taxes as a proportion of National Product more than doubled in the 30 years following 1870, and then doubled again. In the meantime, the progressivity of the tax system also increased (Lindert, 2004). Finally, there is also a consensus amongst economic historians that inequality in Britain fell after the 1870's (see Lindert, 2000, 2004).

Meanwhile, the education system, which was either primarily for the elite or run by religious denominations during most of the nineteenth century, was opened up to the masses; the Education Act of 1870 committed the government to the systematic provision of universal education for the first time, and this was made free in 1891. The school leaving age was set at 11 in 1893, then in 1899, it increased to 12 and special provisions for the children of needy families were introduced (Mitch, 1993). As a result of these changes, the proportion of 10-year olds enrolled in school that stood at 40 percent in 1870 increased to 100 percent in 1900 (Ringer, 1979, p. 207). Finally, a further act in 1902 led to a large expansion in the resources for schools and introduced the grammar schools which subsequently became the foundation of secondary education in Britain.

Following the Great War, the Representation of the People Act of 1918 gave the vote to all adult males over the age of 21, and women over the wage of 30 who were ratepayers or married to ratepayers. Ultimately, all women received the vote on the same terms as men in 1928. The measures of 1918 were negotiated during the war and may reflect to some extent a quid pro quo between the government and the working classes who were needed to fight and produce munitions. Nevertheless, Garrard (2002, p. 69) notes “most assumed that, if the system was to survive and ‘contentment and stability prevail’, universal citizenship could not be denied men, perceived to have suffered so
much and to have noticed Russia’s Revolution.”

Overall, the picture which emerges from British political history is clear. Beginning in 1832, when Britain was governed by the relatively rich, primarily rural aristocracy, a series of strategic concessions were made over an 86 year period to adult men. These concessions were aimed at incorporating the previously disenfranchised into politics since the alternative was seen to be social unrest, chaos and possibly revolution. The concessions were gradual because in 1832, social peace could be purchased by buying off the middle classes. Moreover, the effect of the concessions was diluted by the specific details of political institutions, particularly the continuing unrepresentative nature of the House of Lords. Although challenged during the 1832 reforms, the House of Lords provided an important bulwark for the wealthy against the potential of radical reforms emanating from a democratized House of Commons. Later, as the working classes reorganized through the Chartist movement and later through trade unions, further concessions had to be made. The Great War and the fallout from it sealed the final offer of full democracy. Though the pressure of the disenfranchised played less of a role in some reforms than others, and other factors undoubtedly played a role, the threat of social disorder was the main driving force behind the creation of democracy in Britain.

The story of the rise of mass democracy that emerges from the British evidence is one where economic and social changes connected with industrialization (for example, rising inequality) and urbanization increased the de facto power of the disenfranchised. In response, they demanded political rights, in particular changes in the political institutions which would allocate future political power to them. These changes in political institutions were, in many ways, the direct cause of the changes in economic institutions, in particular, in the labor market, in government policy, in the educational system, with major distributional implications, including the fall in inequality.

Why did elites in Britain create a democracy? Our discussion makes it clear that democracy did not emerge from the voluntary acts of an enlightened elite. Democracy was, in many ways, forced on the elite, because of the threat of revolution. Nevertheless, democratization was not the only potential outcome in the face of pressure from disenfranchised, or even in the face of the threat of revolution. Many other countries faced the same pressures and political elites decided to repress the disenfranchised rather than make concessions to them. This happened with regularity in Europe in the nineteenth century, though by the turn of the twentieth century most had accepted that democracy was inevitable. Repression lasted much longer as the favorite response of elites in Latin America, and it is still the preferred option for current political elites in China or Burma.
The problem with repression is that it is costly. Faced with demands for democracy political elites face a trade-off. If they grant democracy, then they lose power over policy and face the prospect of, possibly radical, redistribution. On the other hand, repression risks destroying assets and wealth. In the urbanized environment of nineteenth century Europe (Britain was 70% urbanized at the time of the Second Reform Act), the disenfranchised masses were relatively well organized and therefore difficult to repress. Moreover, industrialization had led to an economy based on physical, and increasing human, capital. Such assets are easily destroyed by repression and conflict, making repression an increasingly costly option for elites. In contrast, in predominantly agrarian societies like many parts of Latin America earlier in the century or current-day Burma, physical and human are relatively unimportant and repression is easier and cheaper. Moreover, not only is repression cheaper in such environments, democracy is potentially much worse for the elites because of the prospect of radical land reform. Since physical capital is much harder to redistribute, elites in Western Europe found the prospect of democracy much less threatening.

Faced with the threat of revolt and social chaos, political elites may also attempt to avoid giving away their political power by making concessions, such as income redistribution or other pro-poor policies. The problem with concessions however is their credibility, particularly when de facto power is transitory. For example, if a crisis, such as a harvest failure or business cycle recession creates a window of opportunity to solve the collective action problem and challenge the existing regime, the elites would like to respond with the promise of concessions. Yet windows of opportunity disappear and it is difficult to sustain collective action which entails people protesting in the streets and being away from their families and jobs. Thus collective action quickly dissipates and once it does so, the government has an incentive to renege on its promise of concessions. The promise of concessions, which people know to be non-credible is unlikely to defuse collective action. Hence, Acemoglu and Robinson (2000a, 2001, 2003) argue that democratization occurred as a way of making credible commitments to the disenfranchised. Democratization was a credible commitment to future redistribution, because it reallocated de jure political power away from the elites to the masses. In democracy, the poorer segments of the society would be more powerful and could vote, in other words, could use their de jure political power, to implement economic institutions and policies consistent with their interests. Therefore, democratization was a way of transforming the transitory de facto power of the disenfranchised poor into more durable de jure political power.
9.4 Summary

The emergence of mass democracy is another example illustrating our theory of institutions. Into the nineteenth century, economic institutions, particularly in the labor market, disadvantaged the poor. For example, trade unions were illegal and as late as the 1850s in Britain workers trying to organize a union could be shipped to the penal colony in Tasmania, Australia. The poor could not alter economic institutions in their favor because, being disenfranchised, they had little de jure political power and also limited de facto power, the because they were often unable to solve their collective action problems.

However, changes in the structure of society and the economy during the early nineteenth century altered the balance of political power, in particular making the exercise of de facto power by the politically disenfranchised much easier (Tilly, 1995, and Tarrow, 1998, document the changing qualitative nature of collective action over this period). The rise in the de facto political power of the poor necessitated a change in political institutions in their favor to defuse the threat of revolution. This was to tilt the future allocation of de jure political power, and consequently to ensure future economic institutions and policies consistent with their interests.

Whether or not increases in de facto power translated into democracy depended on a number of factors, in particular how difficult and costly it was for elites to use repression to counter the increase in the power of the masses, and how costly the prospect of democracy was. The changes in political institutions that occurred with democracy had profound implications for economic institutions. In the case of Britain, the period after the Second Reform Act of 1867 led the British state to commit itself to providing universal education for the first time and it also led to radical changes in labor market institutions allowing trade unions to form legally for the first time and increasing the bargaining power of labor. Hence economic institutions changed radically in favor of those newly endowed with de jure political power, mostly the relatively poor. This is in fact a relatively general result of democratization. Democracy enfranchises the poor, and the poor are able to use democracy to tilt economic institutions and the distribution of income in society in their favor (Li, Squire and Zou, 1998, Rodrik, 1999).

The emergence of democracy in the nineteenth-century Europe therefore also illustrates the workings of our theoretical framework. In particular, it shows how political institutions determine economic institutions and policies, and thus the distribution of resources, and it shows how political institutions change, especially in response to an im-
balance of de facto political power, as a credible way of influencing the future allocation of de jure political power.

10 Future Avenues

In this chapter we have proposed a framework for thinking about why some countries grow faster and are richer than others. We emphasized, following North and Thomas (1973), that most economic growth theory focuses only on proximate determinants. Although this body of work has been useful in helping us understand the mechanics of growth, it fails to provide a satisfactory account of why some countries grow while others do not. A major research goal must now be to get beyond the neoclassical growth model and its extensions, and search for the deeper causes, i.e., the fundamental determinants of growth.

We argued that the available evidence is consistent with the view that whether or not a society grows depends on how its economy is organized—on its economic institutions. We then proposed the outlines of a theory of institutions and illustrated it through a series of historical examples. We emphasized that a theory of why different countries have different economic institutions must be based on politics, on the structure of political power, and the nature of political institutions. Much remains to be done. First, the framework we outlined was largely verbal rather than mathematical, and thus, by its very nature, not fully specified. Constructing formal models incorporating and extending these ideas is the most important task ahead. Although some of our past work (e.g., Acemoglu and Robinson, 2000a, 2001, Acemoglu 2003b) formalizes parts of this framework, the full model has not been developed yet.

There are also many important issues left out of our framework, which appear to offer fruitful areas for future research. First, though we know that institutions, both economic and political, persist for long periods of time, often centuries (and sometimes millennia), we do not as yet have a satisfactory understanding of the mechanisms through which institutions persist.

Second, and closely related, although institutions do generally persist, sometimes they change. We have important examples of societies which have radically changed their political and economic institutions. Some do so for internal reasons, such as France after the Revolution of 1789, and some do because of external pressures such as Japan after the Meiji restoration or Russia after the Crimean War.

The important point here is that both institutional persistence and institutional change are equilibrium outcomes. Approaches positing institutional persistence as a
matter of fact, and then thinking of institutional changes as unusual events will not be satisfactory. Both phenomena have to be analyzed as part of the same dynamic equilibrium framework.

One type of institutional change, consistent with the examples we discussed in this chapter, takes place when those who benefit from the existing set of institutions are forced to accept change, either because they are the losers in a process of fighting or because of the threat of internal revolution (another possibility is that they might accept change because of the threat of external invasion). However, institutional change can also take place because the set of economic institutions that is optimal for a particular group with political power may vary over time as the state variables in the system and economic opportunities evolve. One example may be the end of slavery in the British Empire and another may be the economic and political changes introduced by Mikhail Gorbachev in the Soviet Union in the 1980s. We need more research on the dynamic mechanisms at work.

Finally, it is important to understand the role of policy and interventions in changing the institutional equilibrium. Though social science research is of intrinsic interest, one would hope that a convincing fundamental theory of comparative growth based on institutions would lead to policy conclusions that would help us improve the institutions and thus the lives and welfare of people in poor countries. It should be obvious that, at the moment, we are a long way from being in a position to draw such conclusions. In a world where political choices are made rationally and are endogenous to the structure of institutions, which are themselves ultimately endogenous, giving policy advice is a conceptually complex issue (see Acemoglu, Johnson, Robinson and Thaicharoen, 2003, for reflections on this issue). Recognizing our current ignorance on this topic in no way diminishes its importance, and its role as the Holy Grail of political economy research, however. And we believe that better and empirically more realistic theoretical frameworks in the future will take us closer to this Holy Grail.


in Political Perspective,” NBER Working Paper #8831


Economics, 3, 1-44.


Davis, Ralph (1973b) English Overseas Trade 1500-1700, Macmillan; London.


Gutierrez, Hector (1986) “La Mortalite des Eveques Latino-Americains aux XVIIe


Murphy, Kevin J., Andrei Shleifer and Robert W. Vishny (1989b) "Income

**Myrdal, Gunnar** (1968) *Asian Drama; An Inquiry into the Poverty of Nations*, 3 Volumes, Twentieth Century Fund, New York.


Historical Review, 103, 705-736.


**Pirenne, Henri (1937)** *Economic and Social History of Medieval Europe*, New York; Harcourt, Brace and Company.


**Ringer, Fritz (1979)** *Education and Society in Modern Europe*, Bloomington; University of Indiana Press.


**Rosenstein-Rodan, Paul (1943)** “Problems of Industrialization in Eastern and


Schattschneider, Elmer E. (1935) *Politics, pressures and the tariff; a study of free private enterprise in pressure politics, as shown in the 1929-1930 revision of the tariff*, New York, Prentice-Hall, inc.


Figure 1
Average Protection Against Risk of Expropriation 1985-95 and log GDP per capita 1995
Figure 2
Latitude and log GDP per capita 1995
Figure 3
GDP per capita in North and South Korea, 1950-98
Figure 4
Urbanization in 1995 and log GDP per capita in 1995
Figure 5
Urbanization in 1500 and log GDP per capita in 1995, among former European colonies
Figure 6
Log population density in 1500 and log GDP per capita in 1995, among former European colonies.
Figure 7
Urbanization in 1000 and 1500, among non-colonies
Figure 8
Urbanization in 1000 and 1500, among former European colonies
Figure 9
Urbanization in 1500 and log GDP per capita in 1995, among non-colonies
Figure 10
Evolution of urbanization among former European colonies

Urbanization in excolonies with low and high urbanization in 1500
(averages weighted within each group by population in 1500)
Figure 11
Evolution of industrial production per capita among former European colonies

Industrial Production Per Capita, UK in 1900 = 100
(from Bairoch)
Figure 12
Urbanization in 1500 and average protection against risk of expropriation 1985-95
Figure 13
Log population density in 1500 and average protection against risk of expropriation 1985-95
Figure 14
Log mortality of potential European settlers and average protection against risk of expropriation 1985-95
Figure 15
Log mortality of potential European settlers and log GDP per capita in 1995
Figure 16
Log population density in 1500 and log GDP per capita in 1995, among former British colonies
Figure 17

Log population density in 1500 and log GDP per capita in 1995, among former European colonies with current population less than 5% of European descent.
The Hyperglobalization of Trade and Its Future

Arvind Subramanian and Martin Kessler

Abstract

This paper describes seven salient features of trade integration in the 21st century: Trade integration has been more rapid than ever (hyperglobalization); it is dematerialized, with the growing importance of services trade; it is democratic, because openness has been embraced widely; it is criss-crossing because similar goods and investment flows now go from South to North as well as the reverse; it has witnessed the emergence of a mega-trader (China), the first since Imperial Britain; it has involved the proliferation of regional and preferential trade agreements and is on the cusp of mega-regionalism as the world’s largest traders pursue such agreements with each other; and it is impeded by the continued existence of high barriers to trade in services. Going forward, the trading system will have to tackle three fundamental challenges: In developed countries, the domestic support for globalization needs to be sustained in the face of economic weakness and the reduced ability to maintain social insurance mechanisms. Second, China has become the world’s largest trader and a major beneficiary of the current rules of the game. It will be called upon to shoulder more of the responsibilities of maintaining an open system. The third challenge will be to prevent the rise of mega-regionalism from leading to discrimination and becoming a source of trade conflicts. We suggest a way forward—including new areas of cooperation such as taxes—to maintain the open multilateral trading system and ensure that it benefits all countries.

JEL Codes: F42, F15, F60, F68

Keywords: Globalization, Convergence, Inequalities, Multilateral Trading System, China

Arvind Subramanian is a senior fellow at the Peterson Institute for International Economics and the Center for Global Development. Martin Kessler is a research analyst at the Peterson Institute for International Economics.

Note: This paper is a modified version of a chapter prepared for the Towards a Better Global Economy Project supported by the Global Citizen Foundation. It is available at http://www.gcf.ch/?page_id=1272. The authors are grateful to Richard Baldwin, Nancy Birdsall, Bill Cline, Kemal Derviş, Shahrokh Fardoust, Theo Priovolos, Dani Rodrik, Andrew Steer, and their colleagues at the Peterson Institute for International Economics and the Center for Global Development for helpful discussions. This paper draws on work with Aaditya Mattoo and on Subramanian (2011). The authors alone are responsible for its content.
1. INTRODUCTION

The post-World War II period witnessed a rapid rise in trade between nations, reminiscent of the integration that occurred before World War I (see WTO 2013 and Krugman 1995). This evolution was facilitated partly by reductions in policy barriers—first in the advanced economies, under the auspices of the then General Agreement on Tariffs and Trade (GATT), and later in developing countries, through unilateral liberalization actions or under programs with the International Monetary Fund (IMF) and World Bank. Trade was also facilitated by technological advances, especially in shipping and transportation. By the end of the 1980s and early 1990s, global trade integration had reverted to levels last seen before World War I.

The postwar period also saw a number of growth successes, beginning with Japan (and Europe), followed by the East Asian tigers and then China, and more recently by India. Along the way, a few countries in sub-Saharan Africa and Latin America also succeeded in raising their standards of living.

In the late 1990s, however, a striking change occurred in the economic fortunes of countries: Economic growth took off across the world, a phenomenon that is best described as convergence with a vengeance. Until the late 1990s, only about 30 percent of the developing world (21 of 72 countries) was catching up with the economic frontier (the United States), and the rate of catch-up was about 1.5 percent per capita per year (table 1.1).1 Since the late 1990s, nearly three-quarters of the developing world (75 of 103 countries) started catching up, at an accelerated annual pace of about 3.3 percent per capita. Although developing country growth slowed during the global financial crisis (2008–12), the rate of catch-up with advanced countries was not materially affected and remained close to 3 percent.

At around the same time, perhaps just preceding this convergence phase, world trade surged, ushering in an era of hyperglobalization. That rising globalization (hereafter used interchangeably with trade integration) is associated with stronger growth, and is a prerequisite for improving the situation of average citizens all over the world, which is reason enough to seek to sustain it. This integration need not continue at the torrid pace of recent years; it should be sustained at a relatively steady rate and any serious reversal, which could set back the prospects of the average global citizen, must be avoided.

This paper is divided into six sections. The next section documents some of the salient features of this era of hyperglobalization. Section 3 discusses three key areas where the trading system is seen as inadequate. The problems are illustrative of the proximate challenges and possible solutions, but in important ways they cannot be solved unless the more fundamental challenges of globalization are addressed. Section 4 explores these deeper challenges. Section 5 suggests possible policy responses at the

---

1. All growth figures in this paragraph use a GDP measure in purchasing power parity terms.
national and international levels that could help sustain globalization. Section 6 offers brief concluding remarks.

The paper is not comprehensive: It focuses on the trade aspects of globalization. It does not discuss other important forms of globalization relating to the movement of finance and people. Rather, it focuses on the major challenges, emphasizing aspects and arguments that have perhaps received less attention thus far.²

2. SEVEN IMPORTANT CHARACTERISTICS OF THE MOST RECENT WAVE OF GLOBALIZATION

This section describes seven major features of the current era of hyperglobalization and of today’s trading system:

- hyperglobalization (the rapid rise in trade integration)
- the dematerialization of globalization (the importance of services)
- democratic globalization (the widespread embrace of openness)
- criss-crossing globalization (the similarity of North-to-South trade and investment flows with flows in the other direction)
- the rise of a mega-trader (China), the first since Imperial Britain
- the proliferation of regional trade agreements and the imminence of mega-regional ones
- the decline of barriers to trade in goods but the continued existence of high barriers to trade in services

Hyperglobalization

Since the early 1990s, the world has entered into an era of what might be called hyperglobalization (figure 2.1). The years between 1870 and 1914 have been described as the first golden age of globalization. World trade as a share of GDP surged from 9 percent in 1870 to 16 percent on the eve of World War I. This was the era that John Maynard Keynes waxed eloquently about, noting that an inhabitant of London “could order by telephone, sipping his morning tea in bed, the various products of the whole earth, in such quantity as he might see fit, and reasonably expect their early delivery upon his doorstep” (Keynes 1920, 11).

The period between 1914 and the end of World War II witnessed the great reversal of globalization, as the combustible mix of isolationism, nationalism, and militarism ignited protectionist policies. World

² For this reason, notable features such as the decline in transportation costs and improvements in information and communication technologies, which have been widely noted, are not studied in depth here (for discussions of these issues, see WTO 2013).
trade plunged to a low of 5.5 percent of world GDP just before World War II began (O’Rourke and Williamson 1999, Frieden 2006, and Irwin 2011).

A third era, starting after World War II, saw the restoration of world trade, aided by declines in transportation costs and trade barriers. Only by about the mid- to late-1970s did world trade revert to the peaks seen before World War I.

The world is now in a fourth era—of hyperglobalization—in which world trade has soared much more rapidly than world GDP. Merchandise exports-to-GDP ratios soared from 15 percent to 26 percent, and goods and services exports to about 33 percent, over the course of the last two decades. This rapid increase is somewhat surprising, because transportation costs do not appear to have declined as rapidly as in earlier eras (Hummels, Ishii, and Yi 2001 and Baldwin 2011a). The cost of information and communications did decline significantly, however.

Part of the increase in trade reflects the fragmentation of manufacturing across borders—the famous slicing up of the value-added chain—as individual production stages are located where the costs of production are lowest. This phenomenon, whereby technology no longer requires that successive stages of manufacturing production be physically contiguous or proximate, has been dubbed the “second unbundling” (Baldwin 2011a).

This real technological impetus to trade tends to artificially inflate recorded trade. Because value is added at each stage of the production chain, it is recorded as exports at successive links in the chain. Gross export flows therefore overstate real flows of valued added (exports net of imported intermediate goods). Figure 2.1 shows that, even though value added-based exports of goods and services are about 5 percentage points lower than exports measured on a gross basis, their trajectory (i.e., for total trade in goods and services) has been similar to that of conventionally measured exports. More recently, value added as a share of exports has not declined substantially or across all trading regions (Hanson 2012 and WTO 2013).

A related feature of this era of hyperglobalization is the rise of multinational corporations and the sharp surge in flows of foreign direct investment (FDI), which have both caused and been caused by cross-border and other flows of goods and services. Since the early 1990s (broadly coinciding with the era

---

3. Throughout this paper, we use trade data as currently measured, on a gross basis. Wherever possible, and as a cross-check, we also present results for trade data measured on a value-added basis. Appendix A explains how these values are calculated.

4. The first unbundling reflected in the quotation from Keynes is the separation of the producer from the consumer that increased trade permits.

5. Koopman, Wang, and Wei (2013) further refined the measurement of value-added trade by distinguishing where countries are (upstream versus downstream) in the value-added chain. The aggregate value-added measures reported here are computed as in their paper.
of hyperglobalization), FDI flows have surged, growing substantially faster than GDP (figure 2.2). Global FDI as a share of world GDP, which hovered around 0.5 percent, increased sevenfold, peaking at close to 4 percent just before the onset of the recent global financial crisis. Even discounting the two surges of 1997–2000 and 2005–08, the general trend is steadily increasing. Global FDI stocks (which are less volatile than flows) jumped from less than 10 percent of GDP in the early 1990s to 30 percent in 2011. FDI flows, and stocks, now surpass levels achieved in the first golden era of globalization, before World War I. By 2009, there were more than 80,000 multinationals, accounting for about two-thirds of world trade (UNCTAD 2010).

**Dematerializing Globalization**

The rapid increase in trade has occurred in both goods and services. Based on conventional (gross) trade data, services trade represented about 17 percent of world trade in 1980 and about 20 percent in 2008. Measured in value-added terms, the corresponding numbers are 30 percent and 43 percent. The apparent paradox that we seek to explain in this section is that services trade, which represents 6 percent of world GDP on a gross basis, is actually 40 to 50 percent larger when computed in value-added terms. This phenomenon arises because services are not always directly tradable but are embodied in the production of goods that are traded. In traditional trade statistics, such a service is not counted as traded, whereas in value-added terms it is considered as such, since the production of this service took place in one country while the service was consumed in another country. Traditional measures of services trade underestimate its importance in global trade.

Two underlying factors can explain the “dematerialization” of trade. First, as Johnson and Noguera (2012) show, the ratio of value-added exports to gross exports in manufacturing goods decreased in the last 30 years (from 60 percent in 1980 to 45 percent in 2009), as a result of the rising importance of global value chains in this sector. Second, as explained in the previous paragraph, trade in services is larger and growing faster in value-added terms than traditional statistics show. Soon, trade in services could eclipse trade in goods, less because services are traded directly, but more because services are increasingly embodied in goods. Trade will actually be dematerializing—moving from “stuff” to “fluff” (intangibles—although the manifestation will be, and the data will record, the opposite effect.

Value added-based trade data reveal how much of total value added in a sector is traded globally. Table 2.1 shows world exports (gross and value added) of goods as a share of world value added in goods (defined to include agriculture and industry), as well as similar numbers for services. During the period of hyperglobalization, value-added exports of goods as a share of total value added in the sector (agriculture and industry) increased from about 33 percent to 47 percent, and services as a share of value added in the services sector increased from 11 percent to 16 percent. Thus, the pace at which services is becoming tradable mirrors that in merchandise.
The slower rise in the tradability of goods than services in the era of hyperglobalization may partly reflect the differential rise in the costs of transportation versus information and communications technologies (ICT). After plummeting sharply between about 1940 and 1980, transportation costs appear to have stabilized (Baldwin 2011a and Hummels 2007). In contrast, after 1990, the use of ICT-related technologies and applications surged. A consequence could have been a differential fillip to more sophisticated goods and especially services.

**Democratic Globalization**

Part of the increase in trade also reflects convergence and the wider distribution of output and income: That is, trade has grown because output has become more widespread and “democratic.” Basic gravity theory implies that smaller countries tend to trade more than larger ones. A world made up of two equal-size countries will experience more trade than a world in which the larger country accounts for 95 percent of world output. Over time, the world is becoming less unequal in terms of the distribution of the underlying output that generates trade. For example, between 1970 and 2000 the world was constituted by about 7.0–7.5 country equivalents (with fluctuations) (figure 2.4). Since 2000, as more countries have started catching up with the rich, world output has become more dispersed: Today, it is as if there were 10 country-equivalents in the world. In the era of hyperglobalization, roughly a third of the increase in trade can be accounted for by this democratization of world output (figure 2.3).

Even if the rise in world trade is caused by spreading prosperity, is this rise itself broadly spread? The numbers in figure 2.1 are in effect a GDP-weighted average of individual country’s export-to-GDP ratios. We can, instead, calculate export-to-GDP ratios that are unweighted or weighted by population to measure the reach of globalization across countries and across people, as done in figure 2.4.

Figure 2.4 shows that in 1913, the peak of the first golden era of globalization, the unweighted average export-to-GDP ratio in the world was close to 15 percent. In 2010, it was 21.5 percent. The population-weighted export-to-GDP ratio was about 6 percent; by 2010, it was more than 15 percent. Hyperglobalization has thus come about not just because some rich countries are becoming more open

---

6. The gravity model of trade is theoretically well-established and empirically validated. It shows that trade between two countries is proportional to their economic size and inversely proportional to their distance. Other things equal, a large country will trade more than a small one but will be less open (trade/GDP will be smaller).

7. As Anderson (2011) shows, in a world without trade frictions, the share of trade in world output is given by \( 1 - \sum b_j^2 \) where \( b \) is the share of a country in world output. Inverting the expression gives the number of country-equivalents in the world, which increases with convergence. Baier and Bergstrand (2001) find a statistically significant effect of convergence on trade.
but also because openness is being embraced more widely. Keynes’ paean to globalization was thus both imperialist and elitist.

Criss-Crossing Globalization

Trade has been increasing steadily. But one of the unique features of the most recent phase of hyperglobalization is the fact that similar kinds of goods (and capital) are criss-crossing global borders. In other words, it is less and less the case that a country’s imports and exports are very different.

Three manifestations of such criss-crossing globalization can be discerned. In the immediate aftermath of World War II, the industrial countries increasingly started to export and import manufactured goods (for example, Japan, Germany, and the United States all exported and imported cars), a phenomenon at odds with classic Ricardian model. Models of monopolistic competition (Helpman and Krugman 1985) combined with consumers’ love for variety (differentiated products) provided the theoretical basis for the phenomenon of intraindustry trade that related to trade in final goods. Melitz and Trefler (2012) show that the share of intraindustry trade in total trade increased by nearly 20 percentage points. But this increase occurred between 1960 and the mid-1990s rather than over the most recent period of hyperglobalization. In fact, since the 1990s, this share of intraindustry trade has stabilized (Brülhart 2009).

For the rapidly growing emerging market countries of Asia, criss-crossing globalization has taken the form of greater two-way flows of parts and components than of final goods. This phenomenon is related to the slicing up of the value-added chain and the unbundling noted above.

The share of parts and components in trade offers one measure of criss-crossing globalization. For the world as a whole, this share increased from about 22 percent in 1980 to 29 percent in 2000. Since then, intermediate goods trade declined to about 26 percent of total trade, suggesting that the internationalization of production may have peaked (WTO 2013). Indeed, this form of globalization was really observed only in Asia, and even there intermediate trade has declined since 2000. Even in China, reliance on imports has declined markedly. In the computer sector (broadly defined), for example, exports

8. One potential problem with figure 2.4 is that the sample is not constant over time. The finding that trade has become more democratized holds even for the constant sample of countries (not reported here). The unweighted average is above the population-weighted average because populous countries tend to trade less.

9. Even within the United Kingdom, the benefits of globalization were not broadly accessible. In 1912, for example, there were 0.6 million telephone subscribers in the United Kingdom, the population of which was about 46 million.

10. Another way of describing this democratization is to note that the trade of low- and middle-income countries has grown more rapidly than their incomes and more rapidly than the trade of high-income countries and that a bulk of this growth is trade among low- and middle-income countries (Hanson 2012).
were only 1.6 times imports in 1994, indicating substantial intermediate trade; by 2008, this ratio had climbed to 4.2 (Hanson 2012).

The third (and perhaps least remarked on) dimension of criss-crossing globalization, with potentially important effects for globalization policies, relates to two-way flows of FDI. It is one of the unique aspects of this era of hyperglobalization that developing countries (especially the larger ones) are exporting FDI (which embodies sophisticated factors of production, including entrepreneurial and managerial skills and technology)—and not just to other developing or countries (Mattoo and Subramanian 2010). Figure 2.5 plots a Grubel-Lloyd index of two-way flows of FDI at the global level. Depending on the weighting, this index climbed from about 0.3 in 1970 to almost 0.7 by 2011.

The Rise of a Genuine Mega-Trader: China

When Krugman (1995) surveyed the evolution of world trade, he noted as one of the distinctive features the rise of a number of Asian super-traders, including Singapore, Hong Kong (China), and Malaysia, all of whose exports exceeded 50 percent of GDP, a feature never seen in the first era of globalization (in 1913, the United Kingdom’s ratio of export to GDP was 18.5 percent). But mega-traders can be defined in two senses: globally (relative to world trade) and nationally (relative to a country’s own output). Krugman clearly applied the latter criterion. Had he applied the former, one mega-trader he would have identified would have been Japan in the 1980s, which accounted for about 7.5 percent of global trade at its peak. Based on this criterion, none of the other East Asian Tigers would have been noteworthy, despite their astonishing performance: The small economies of Singapore, Hong Kong (China), Taiwan (China), and Malaysia accounted for a very small share of world trade at their peaks.

Since 1990, a true mega-trader has emerged: China. It qualifies as such under both definitions of the term. Its integration to world trade has accelerated with its accession to the WTO in 2001, and transformed the country into the world’s largest exporter and importer of manufactured goods, having surpassed the United States in 2012 (table 2.2).

China’s exports as a share of GDP are now almost 50 percent. When its size and income level are taken into account, it is a substantial over-trader, comparable to the United Kingdom in the heyday of its empire and a vastly bigger trader than the United States, Japan, or Germany at their peaks.

For example, in 1975, the United States’ trade-to-GDP ratio was 16.1 percent (table 2.3). Given the size and income level of the United States, that number represented under-trading of about 35 percent. Japan in 1990, with a trade-to-GDP ratio of 20 percent, under-traded by about 50 percent. In

11. The Grubel-Lloyd index, which can take values between 0 and 1, measures the degree of two-way flows for a given country or industry. An index of 0 denotes that a country’s exports and imports are perfectly dissimilar—that is, a country is either fully an importer or an exporter of a good (or, in this specific case, a type of capital flow). An index of 1 denotes that a country’s exports and imports are similar—that is, a country exports and imports of a certain good are identical in magnitude.
contrast, China’s trade-to-GDP ratio in 2008 was 62.2 percent, which represented over-trading of nearly 60 percent. Only Imperial Britain was a mega-trader in both senses of the term. In 1913, its exports represented 18.5 percent of world exports. Its export-to-GDP ratio was 12 percent, which represented over-trading of about 84 percent. China is thus the first mega-trader since Imperial Britain.

If trade continues to grow in line with income, China’s dominance in world trade will become even greater. According to simple calculations in Subramanian (2011), by 2030 China could account for about 16 to 17 percent of world exports, nearly three times the share of the United States (see table 2.2). Even at the height of US dominance, around 1975, it did not account for as large a share of world trade or have as great an edge over its nearest competitors (in 2000, the United States accounted for about 16 percent of world exports compared with 8 percent for Germany and about 7 percent for Japan). Any discussion of trade and the trading system going forward must recognize this development (discussed further below).

**Growing Regionalization, Preferential Trade, and Impending Hyperregionalization**

The era of hyperglobalization has been accompanied by a proliferation of preferential trade agreements (PTAs). Today, about half of the exports of the top 30 exporters go to preferential trade partners. Between 1990 and 2010, the number of PTAs increased from 70 to 300 (figure 2.6). In the mid-1990s, about 75 percent of PTAs were regional; by 2003, this share had dropped to about 50 percent. All World Trade Organization (WTO) members except Mongolia have concluded at least one PTA; some, such as the European Union, Chile, and Mexico, have concluded more than 20. Some of the large traders have already concluded agreements with each other or are about to do so (examples include the European Union and Mercosul, Japan and Mercosul, the European Union and India, and India and Japan).

The fact that nearly half of world trade is covered by preferential agreements does not mean that a comparable figure enjoys preferential barrier reductions. Carpenter and Lendle (2010) estimates that only about 17 percent of world trade is eligible for preferences; the remaining 83 percent either enjoy zero nondiscriminatory tariffs (nearly 50 percent) or are excluded from preferential agreements. Moreover, where preferences can apply, margins are low. For example, less than 2 percent of world imports enjoy preferences greater than 10 percentage points.

An interesting new dimension of these PTAs is the extent to which they feature “deep integration” (Lawrence 1996)—that is, liberalize not only tariffs and quotas but other “behind-the-border” barriers, such as regulations and standards, as well. In the last 10 years, for example, nearly 40 agreements have included provisions on WTO+ issues (competition policy, intellectual property rights, investment, and the movement of capital). This figure is four to five times greater than comparable agreements in the pre-WTO era (WTO 2011) (see figure 2.6 and table 2.4).

---

12. The WTO (2013) projection for 2035 is exactly in line with the estimate in Subramanian (2011). The WTO’s mean estimate is that China will account for 17 percent of world trade in 2035, with a range of 11 to 23 percent.
In part because of these deep integration agreements, it would be wrong on the basis of the tariff evidence to underestimate the potential discriminatory effect of preferential arrangements. In agriculture and some manufacturing sectors, such as textiles, tariffs are still high. In services, any future deepening of preferential agreements could create significant discrimination against outsiders, because most favored nation (MFN) levels of protection are significant and there is considerable scope for the preferential recognition of standards, licensing, and qualification requirements (Chen and Mattoo 2004). Strong exclusionary effects could also arise from “deeper integration” along other dimensions: Preferential agreements increasingly have provisions on investment protection, intellectual property rights, government procurement, competition policy, and technical barriers to trade.

On regional agreements, seismic changes are under way, with the possible negotiation of mega-regional agreements between the United States and Asia (the Trans-Pacific Partnership) and the United States and Europe (the Transatlantic Trade and Investment Partnership). Trade between these groups of countries accounts for about $2 trillion to $3 trillion a year in world trade, signifying a potentially major jump in the volume of trade covered by preferential agreements. These PTAs would represent the first between the top four major regions of the world (China, the United States, Europe, and Japan), with consequences that will be discussed below. If the Transatlantic Trade and Investment Partnership and Trans-Pacific Partnership (to the extent that it includes Japan) are concluded, about 20 percent of global trade will be covered by those deeper regional agreements. It is not unforeseeable to think of an era in which nearly all trade becomes regional.

**Lower Formal Barriers in Goods, High Barriers in Services**

The world has become much less protectionist. Globally, MFN tariffs have declined from more than 25 percent in the mid-1980s to about 8 percent today. Border barriers (tariffs and nontariff measures) in manufacturing in the Organization for Economic Cooperation and Development (OECD) countries are less than 4 percent. The US International Trade Commission (USITC 2011) estimates that the welfare gains in the United States from eliminating all remaining tariffs are close to zero. Border barriers in the larger emerging markets are higher, but they have declined considerably, from about 45 percent in the early 1980s to just over 10 percent in 2009 (figure 2.7).

But barriers to trade in services remain high. Borchert, Gootiz, and Mattoo (2012) calculate trade restrictiveness indexes for services. They cover five major sectors—financial services, telecommunications, retail distribution, transportation, and professional services—and the different modes of delivering these services across borders and via investment abroad). The index ranges from 0 (completely free) to 100 (completely restricted).

13. There are no data on barriers to trade in services going back in time that would allow a quantitative description of changes in barriers.
Barriers vary across service sectors, but the average level is high (figure 2.8). Barriers are relatively low in telecommunications and relatively high in transportation and professional services. They also vary across regions: Latin America is nearly as open as OECD countries, whereas Asia and the Middle East have high barriers. In fact, as in goods, barriers are correlated with a country’s level of development (figure 2.1). What this means is that international negotiations will increasingly focus on services and FDI.

Two points are worth noting. First, barriers to trade in goods and services have declined sharply over time; the world as a whole is thus becoming less closed. But the composition of world trade is shifting toward the poorer countries (especially toward China and India), and these countries are on average more protectionist (as figures 2.7 through 2.9 illustrate). The composition of world output is also shifting toward services and away from manufacturing. Both these compositional shifts make the world as a whole less open and attenuate the liberalization trend that stems from all countries reducing barriers.

Second, the integration of goods and services markets is nowhere close to completion. One way of assessing how far from full globalization the world still is might be to compare actual trade with what is predicted by a simple gravity model without frictions. As Krugman (1995) and Anderson (2011) show, under frictionless trade, the world trade share is inversely related to the distribution of GDP across countries: the more equal the distribution, the greater the world trade share. In 1970, actual trade was about 10 percent of the theoretical maximum predicted by the frictionless gravity model. In 2011, it was about 40 percent (perhaps less if trade is calculated on a value-added basis). Thus, although actual trade is rapidly catching up with trade in a frictionless world, there is still some way to go.

3. THREE PRESSING PROXIMATE CHALLENGES

This section discusses three recent challenges that have emerged in the trading system and proposes potential solutions to each of them. The proposed solutions can never be reached on their own, however, unless the deeper and more fundamental challenges, discussed in subsequent sections, are addressed.

Trade and Currency Wars

Mercantilism and Self-Insurance: The Dual Origins of Reserve Accumulation

In the late 1990s, in the aftermath of the Asian financial crisis, a number of emerging market countries, especially in Asia, adopted an economic strategy that was dubbed Bretton Woods II (Dooley, Folkerts-Landau, and Garber 2003). This strategy had two motivations and one manifestation. Reeling from the

---

14. This index cannot strictly be compared with tariffs, but the farther away the number is from zero, the less open a country is.

15. An implication of this finding on unrealized globalization is that going forward, this potential is likely to be greater in services than in goods.

16. The next section, on trade and currency wars, draws on Mattoo and Subramanian (2009).
disruption that sudden withdrawal of foreign capital had caused to their economies and chafing at the political humiliation of having to borrow from the IMF, they decided to self-insure against future crises.\textsuperscript{17} Self-insurance took the form of building an arsenal of foreign exchange reserves (see Allen et al. 2013; Goldstein 2009). The second motivation was mercantilism, a strategy that made trade surplus the engine of growth.

Both motivations translated into a common manifestation: Countries moved from being large net importers of finance (running large current account deficits) to being less reliant on finance, or in some cases, notably China, Malaysia, and Taiwan (China), to becoming net exporters of finance (running current account surpluses). These motivations also translated into—actually require—a policy of undervaluing the exchange rate in a fixed or managed peg regime, aided by intervention in foreign exchange markets. A few countries in East Asia (China and Malaysia in particular) tended to maintain restrictions on capital inflows as a way of sustaining a competitive exchange rate.

Bergsten and Gagnon (2012, 2) argue that more than 20 countries have been intervening in foreign exchange markets for several years “at an average rate of nearly $1 trillion annually… to keep their currencies undervalued and thus boost their international competitiveness and trade surpluses.” These countries include China and a number of East Asian countries, oil exporters, and some advanced countries, including Israel and Switzerland.

This problem is not new. Similar issues of undervaluation arose relating to the Deutsche mark in the 1960s and the yen in the 1970s and 1980s. The mercantilism or currency wars of today are related to the much deeper problem—and some would argue the greatest design flaw in the Bretton Woods system—of the asymmetric adjustment between surplus and deficit countries in the international monetary and trade system. Bergsten and Gagnon (2012, 10) write that “it is a huge irony that the Bretton Woods system was created at the end of the Second World War primarily to avoid repeating the disastrous experiences of the inter-war period with competitive devaluations, which led to currency wars and trade wars that in turn contributed importantly to the Great Depression, but that the system has failed to do so.”

\textit{Consequences of Mercantilism}

Why are current account surpluses combined with undervalued exchange rates a problem for the international economic system? The consequences or problems can be categorized as cyclical mercantilism, structural mercantilism, and macro-mercantilism.

Cyclical mercantilism arises when the economy is depressed relative to trend growth; such a situation is characterized by idle resources, underutilized capital, and unemployment. Mercantilism by

\textsuperscript{17} The perceived humiliation was captured in the picture of the IMF Managing Director at the time, Michel Camdessus, looking over a head-bent President Suharto signing the economic adjustment program with the IMF.
one country threatens deflation in partner countries facing idle resources. This concern preoccupied Keynes, who argued that because of international liquidity constraints, there would always be greater pressure on debtors (countries running deficits) to adjust than on creditors. This asymmetry would impart a deflationary bias, because debtor countries would have to cut demand without surplus countries having to undertake the offsetting reflation. He therefore proposed that pressure be exerted on creditor countries by forcing them to pay instead of receiving interest on their positive balances (see Williamson 2011).

In the current context, the deflationary impact of Chinese policies on the United States, quantified by Cline (2005), has prompted some commentators to call for aggressive trade action by the United States against China and other countries practicing mercantilism (Krugman 2010 and Bergsten and Gagnon 2012). Persistent surpluses by Germany and their deflationary impact, especially within Europe, have sparked similar calls for action (Wolf 2010).

Structural mercantilism arises when a country pursues policies such as undervalued exchange rates as development or growth policy for an extended period of time. Such policies can have long-run effects on partner countries. An undervalued exchange rate is both an import tax and an export subsidy; it can have adverse effects on trading partners. One way structural mercantilism is transmitted is by depressing the medium-run price of manufactured products, reducing opportunities for specialization in manufacturing and manufactured exports in partner countries. The concerns expressed in sub-Saharan Africa and Latin America relate to structural mercantilism. Mattoo, Mishra, and Subramanian (2012) show, for example, that a 10 percent depreciation of China's real exchange rate reduces a developing country's exports of a typical product to third markets by about 1.5 to 2.0 percent. Such a decline can have long-run growth effects.

Macro-mercantilism was most evident in the recent crisis in the creation of the so-called savings glut (Bernanke 2005, 2007). Large and growing aggregate current account surpluses increase global liquidity, leading to easy credit and lending, which can easily morph into imprudence, financial excess, and asset bubbles, threatening financial stability. The savings glut hypothesis is by no means uncontested; many economists argue that monetary and regulatory policies in borrowing countries should bear the brunt of responsibility (Johnson and Kwak 2010 and Haldane 2010). How much blame the bartender should bear for plying alcohol on a drinker who binges will forever be disputed. But that excess liquidity was a factor and that Chinese mercantilist policies created excess liquidity are plausible deeper causes of the Lehman crisis (see Bernanke et al. 2011).

Currency wars or the resulting global imbalances are a systemic problem only if one or a few large countries pursue them. The possibility of collective action to prevent them must take account of this reality.

Exchange rates and foreign exchange intervention are centrally implicated in mercantilism. The international monetary system, under the auspices of the IMF, is therefore the best forum in which to
find a solution. The prospects for any serious reform remain slim, however, because of the inherent limits to international monetary cooperation. Systemic threats arise from the policies of the largest countries, in particular when policies pursued in self-interest conflict with the collective interest. But, by definition, it is difficult for the rest of the world to change the incentives of the large country to give more weight to the collective interest. Successful cooperation is fated to falter if not fail—and the history of the IMF in this matter has often resulted in failure.

As Mussa (2007) makes clear, “In none of these consultations has the Executive Board ever concluded that a member was out of compliance with its obligations regarding its exchange rate policies or any other matter” (emphasis in the original).18

Williamson (2011, 1) notes that “it has been 80 years since John Maynard Keynes first proposed a plan that would have disciplined persistent surplus countries. But the Keynes Plan, like the subsequent Volcker Plan in 1972–74, was defeated by the major surplus country of the day (the United States and Germany, respectively), and today China (not to mention Japan or Germany) exhibits no enthusiasm for new revisions of these ideas.” The question is whether there is anything that the rest of the world could have done—by way of sticks or carrots—to have persuaded the United States in 1944, Germany in 1973, or China in 2007 to change its positions or policies for the collective good.

The IMF’s ineffectiveness is a proximate manifestation of deeper structural causes related to leverage and legitimacy. Although the IMF has been able to effect changes in member country policies in the context of financial arrangements, it has not been influential without the leverage of financing. In its key surveillance function (which involves no financing), the IMF has rarely led to changes in the policies of large creditor countries, even when such policies have had significant spillover effects on countries; it has not been able to persuade large creditor countries to sacrifice domestic objectives for systemic ones. There seems to be an implicit “pact of mutual nonaggression,” to use Mussa’s phrase, in IMF surveillance. Perhaps as a result, the IMF has had a history and tradition of nonadversarial dialogue between its members and has not had to develop a real dispute settlement system.19

18. Keynes himself recognized the asymmetry of IMF leverage between creditor and debtor countries in the discussion in the lead-up to the creation of the IMF.
19. A corollary of the observation that cooperation is least likely where the self-interest of the largest countries are at stake is that the prospects for successful cooperation are greater where these countries are less affected and when the demands on them are minimal. Building global safety nets by providing greater and more expeditious access to crisis financing is one area where the greatest progress has already been made. The IMF’s lending ability tripled after the crisis, and it may increase further. For the large countries, it is both desirable and effective to push for larger safety nets. The costs are relatively small—involving larger financial contributions rather than any major change of domestic policies—and the rewards are great, because the system as a whole is strengthened while the individual clout of the large countries is increased (see Goldstein 2009).
Compounding this problem of limited leverage is the IMF’s eroding legitimacy. Although its role and importance were rehabilitated with the recent global financial crisis, the perception of the IMF as an unreliable interlocutor in emerging market countries—Asia in particular—endures. A good example is the IMF’s new conditionality-lite financing facility, which has few takers because some emerging market countries do not want to be seen as even potential borrowers from the IMF. Indeed, in 2009, a number of emerging market countries—Brazil, Singapore, and the Republic of Korea—preferred to get lines of credit from the US Federal Reserve rather than from the IMF.

The WTO seems to be different on these two counts of legitimacy and leverage, because it works on the basis of the exchange of concessions, which ensures that all players feel that they have derived a fair political “bargain.” Reciprocity ensures political buy-in to cooperation. Periodic negotiations in the GATT/WTO have updated this political contract between countries, redressing some old grievances and papering over others, with the implicit understanding that there will be a future occasion to take up the unsolvable problems of the day.

A consequence of reciprocity and the periodic updating of the political contract to cooperate—and another reason why the WTO works—is that this process creates incentives to adhere to the dispute settlement contract. Dispute settlement by the WTO is effective largely because countries feel that they have previously (and recently) made a reasonably advantageous, fair, and equitable bargain, to which they must adhere. WTO governance works because negotiations to create the rules and agree on liberalization are perceived as fair and broadly equitable in outcome, rendering subsequent compliance possible.20

Trade, Climate Change, and Green Growth

Do the institutions and ideology of globalization come in the way of tackling climate change? In one very important respect, they may.21

Consider two episodes from 2012. In late 2012, the United States and the European Union sanctioned the use of antidumping duties against Chinese exports of solar panels on the grounds that Chinese manufacturers were “dumping” (selling below cost) solar panels manufactured in China. In the presidential debates, President Obama was on the defensive against Mitt Romney, who tried (with some success) to tar him with the “failed industrial policy” brush in relation to government support for clean energy and Solyndra, a producer of solar panels that filed for bankruptcy two years after receiving substantial government loans and guarantees. These examples illustrate how international rules and ideology (which underlie rules) could interfere with efforts to tackle climate change.

20. Experience suggests that the mere prospect of retaliation, as well as the reluctance to be seen as a rule breaker, is sufficient to ensure compliance and that there is rarely need for action.

21. This section is based on Mattoo and Subramanian (2013).
Mattoo and Subramanian (2012b) argue that only radical technological change can reconcile climate change goals with the development and energy aspirations of the bulk of humanity. Technological change requires the deployment of the full range of policy instruments that would raise the price of carbon and provide incentives for research and development of noncarbon-intensive sources of energy and related green technologies. With notable exceptions, countries have shown great reluctance to raise the price of carbon directly.

International rules severely restrict the use of subsidies. Under current WTO rules, domestic subsidies for the development and production of clean energy and related energy technologies are actionable by partner countries if those countries feel that their domestic production or exports are adversely affected. Until 2000, some of these subsidies were deemed nonactionable, but the exemption has not been renewed. Moreover, all forms of export subsidy involving clean energy and/or green technologies are prohibited. These rules are in place because of the ideology that imbues globalization—the notion that subsidies and all forms of industrial policy are dubious.

In relation to climate change, these rules are doubly bad. There is, of course, a logic to curtailing subsidies: Even if they confer domestic benefits, those benefits are outweighed by the damage to partner countries. A multilateral rule to which there is general adherence reduces that damage, potentially leaving countries better off. But in the case of climate change, because spillovers are global, any subsidy that promotes clean energy and development confers a benefit to partner countries. On balance, therefore, rules should err on the side of promoting rather than restricting subsidies.

There is a second, arguably bigger, political economy benefit. Prospects for climate change action in the United States in the form of a carbon tax or cap-and-trade are not bright. President Obama’s grand rhetoric in his 2013 State of the Union speech is unlikely to be matched by bold action because of the lack of bipartisan support in Congress. This state of affairs reflects a combination of factors—climate change denial, the strength of the carbon energy industries, and weak economic prospects. There is probably only one development that could galvanize action in the United States: the threat that China will capture green technology leadership. The United States needs a Sputnik moment of collective alarm at the loss of economic and technological ascendancy.

The problem is that China is currently constrained by WTO rules, as the actions against its firms in 2012 illustrate. 22 China and all countries that are not straitjacketed by the tyranny of the subsidies-

are bad ideology and that have the financial means to do so should be allowed to deploy industrial policy to promote clean energy and green technologies. If doing so leads to a subsidy war because partners feel threatened, that is a war that should be promoted, as it will ignite the race for the development and production of an undersupplied global public good. From this perspective, WTO rules should allow not only domestic but, perhaps also export subsidies under some conditions; current rules circumscribe the use of domestic and prohibit the use of export subsidies.23

Trade and Scarcity of Food and Resources

The 2007 global food crisis was severe.24 According to the World Bank, about 100 million people are estimated to have been thrown back into the ranks of the poor because of increases in the price of food. Riots occurred in a number of countries. The Bank identified 33 countries as especially vulnerable. The poor were especially vulnerable because they spend the largest portions of their income on food. In the United States, the poor spend an estimated 18 percent of their income on food; a similar measure for households earning less than $1 a day is about 72 percent in Peru and South Africa, 66 percent in Indonesia, and 50 percent in Mexico (Banerjee and Duflo 2011).

But pressure on food supplies, and associated high food prices, could be a medium- to long-term reality, because some of the driving factors—rising prosperity in the developing world, which creates more demand, high fuel prices, stagnant agricultural productivity, and climate change-induced pressure on agricultural supplies, including through the depletion of water—could be of a durable nature. These fundamentals are being exacerbated by export restrictions on foodstuffs. According to a World Bank report, in the 2007 crisis, 18 developing countries imposed some form of export restrictions (Zaman et al. 2008). Each country tried to keep domestic supplies high, on the grounds of food security. But as more countries implemented export controls, global supply contracted, pushing prices up and exacerbating global food insecurity. The global rice market was particularly affected by trade restrictions.25

23. Another area in which trade restrictions should be permitted is border tariffs against imports from countries that do not tax carbon in the manner that the importing country does. Such tariffs would help overcome opposition from energy-intensive industries in countries wishing to raise the price of carbon on the grounds that they would be rendered uncompetitive relative to imports from countries that do not tax carbon. A final area in which WTO rules need to be clarified is export restrictions on natural gas, which is becoming an important fuel. The United States currently limits its exports to countries with which it does not have a free trade agreement. If greater global use of natural gas is desirable (because it is cleaner than substitutes such as oil and coal), then restrictions on exports may be deleterious for global energy emissions.

24. This section draws on Mattoo and Subramanian (2012a).

25. Food security goals are best served not by restricting trade but by deploying domestic policy instruments such as targeted safety nets. The existence of such safety nets would dilute the political economy bias in favor of trade interventions.
There are few restrictions on the use of export taxes in the WTO, and its disciplines on export restrictions are incomplete. The GATT does prohibit quantitative restrictions on exports, but temporary restrictions are permitted in order to prevent critical shortages of food or other goods.

This permissiveness on export taxes and restrictions is resulting in the worst of all possible worlds. Under “normal” agricultural conditions, costly taxpayer support reduces imports and encourages production and exports, creating huge distortions. Under abnormal conditions, such as are prevailing now, the opposite occurs: Countries liberalize their imports but prevent exports. What is needed is a system in which both imports and exports remain free to flow in good times and bad. Such a system is especially important if trade is to remain a reliable avenue for food security. If in bad times importing countries are subject to the export-restricting actions of producing countries, they will consider trade an unreliable way of maintaining food security and reconsider how to manage their agriculture. As a result, there will be a greater temptation to move toward more self-reliance as insurance against the bad times.26

The Doha Round of trade negotiations did not address these problems. It was devoted to traditional forms of agricultural protection—trade barriers in the importing countries and subsidies to food production in producing countries—which are now becoming less important as food prices have soared and import barriers declined. The trade agenda needs to be enlarged, so that trade barriers, on both imports and exports, are put on the trade agenda.

Trade policies have also exacerbated the scarcity of nonfood resources. Concerns have already arisen over China’s restriction of exports of rare earth metals, for some of which (for example, scandium and yttrium) it accounts for more than 70 percent of the world’s exports. It also accounts for a large share of exports of other key raw materials, including various forms of bauxite, magnesium, and zinc.27

4. FUNDAMENTAL POLICY CHALLENGES

The period of hyperglobalization has been associated with the most dramatic turnaround in the economic fortunes of developing countries. Regardless of the view one takes about this association, it is safe to say that a broadly open system is good for the world, good for individual countries, and good for average citizens in these countries. Going forward, even if the pace of hyperglobalization slows, the aim of policy

26. Not surprisingly, WTO members that depend heavily on world markets for food (for example, Japan and Switzerland in 2000; the Democratic Republic of Congo, Jordan, and the Republic of Korea in 2001) have pushed for disciplines on export controls and taxes. Recognizing that importers’ concerns about the reliability of supply could inhibit liberalization, some exporting countries have advocated for multilateral restrictions on the right to use export restrictions (examples include the Cairns Group and the United States in 2000 and Japan and Switzerland in 2008, see the International Economic Law and Policy Blog 2008).

27. In an earlier case, a WTO panel ruled against certain export restrictions China had maintained on a number of raw materials, including bauxite, coke, fluorspar, magnesium, and zinc.
at the national and collective level must be to sustain steady and rising globalization and avoid sharp reversals.

The previous section illustrated some of the proximate challenges. They can be addressed only if the deeper challenges are recognized and addressed.

One way of approaching these more fundamental policy challenges is suggested in table 4.1, which helps identify the problems and hence to prioritize the policy response. This schematic can be applied to three broad groups of countries (high, middle, and low income), the challenges and responses for each of which may differ.

What are the really important challenges for the open trading system, and how should they be responded to? If the next couple of decades mimic or come close to mimicking the last two in terms of globalization, success will have been unambiguous. The challenge is thus simply to maintain the status quo and allow the forces that have shaped globalization over the last few decades to play themselves out.

Alternatively, one could argue that globalization needs to advance on a number of different dimensions—because, for example, impediments remain to the prospects of average citizens, especially in low-income countries. The need for further globalization could also stem from the perception that in some respects, the current system is unsustainable, because it is differentially open and the burden of providing open markets is not equally shared, especially by China.

A third logical possibility is that the forces that will push against globalization are, or will become, so strong that a retreat from current globalization is inevitable. The challenge then will be to manage this retreat in a way that minimizes the costs to countries and citizens around the world.

The responses to each of these challenges can occur at the national level, at the international level, or through some combination of national and collective action. The responses to these challenges are discussed below.

The West’s Challenge: Hyperglobalization Meets Economic Decline

The Bad News

Public support for free trade agreements in the United States is at its lowest point since 2006, according to the Pew Center (2010)—and the decline occurred quickly. In 2009, the share of people who supported free trade agreements exceeded the share who opposed it by a margin of 11 percentage points. In 2010, opponents of free trade outnumbered supporters by 8 percentage points. Surprisingly, among Republican-leaning voters, the turnaround was even more dramatic: The margin in 2009 was 7 percentage points in favor of free trade agreements; the margin in 2010 was 26 percentage points against free trade agreements. This weakening collective perception of the benefits of openness is matched, mirrored, or validated by intellectual opinion.
Samuelson (2004) argues that the rise of developing countries such as China and India could compromise living standards in the United States, because as they move up the technology ladder, they provide competition for US exports, reducing their price. Krugman (2008) focuses on the impact of imports from developing countries, particularly China, on the distribution of income in the United States and the wages of less-skilled workers. His conclusion is that “it is likely that the rapid growth of trade since the early 1990s has had significant distributional effects” and that “it is probably true that this increase (in manufactured imports from developing countries)… has been a force for greater inequality in the United States and other developed countries” (Krugman 2008, 134–35).

Blinder (2009) draws attention to the employment and wage consequences of the outsourcing that has been facilitated by technological change and trade in services. He estimates that 22 to 29 percent of all US jobs will be offshored or offshorable within the next decade or two.

Summers (2008a, 2008b) has highlighted the problems stemming from increasing capital mobility. Hypermobile US capital creates a double whammy for American workers. First, as companies flee in search of cheaper labor abroad, American workers become less productive (because they have less capital to work with) and hence receive lower wages; the “exit” option for capital also reduces the incentive to invest in domestic labor. Second, capital mobility impairs the ability of domestic policy to respond to labor’s problem through redistribution because of an erosion in the tax base as countries compete to attract capital by reducing their tax rates.

Spence and Hlatshwayo (2012) argue that almost the entire increase in employment—27.3 million jobs in the United States between 1990 and 2008—was in the nontradable sectors, where productivity growth was much slower than in the manufacturing and tradable sectors, explaining the long-term stagnation of wages in last segment of the workforce.

That a constellation of intellectuals—instinctively cosmopolitan and ideologically liberal—talks like this is an important signal, not least because the objective circumstances have changed. One might call this challenge that of the irresistible force of globalization and hyperglobalization meeting the immovable object of weakening economic and fiscal fortunes in the West.

In the United States, except for a brief spell in the late 1990s, median wages have stagnated for three decades; inequality has been sharply rising, particularly because of rising incomes at the very top of the income spectrum (Piketty and Saez 2003); and mobility has declined (Haskins, Isaac, and Sawhill 2008). Worse, as in all industrial countries, indebtedness has risen (average debt in the G-7 is now about 80 percent of GDP), prospects for medium-term growth in the future are not bright (according to the latest World Economic Outlook forecast), and aging and entitlements add to the serious fiscal pressures looming ahead. These objective conditions are not the most propitious for sustaining globalization.
This structural malaise is captured in the following metaphor that Larry Katz, of Harvard, uses: “Think of the American economy as a large apartment block. A century ago—even 30 years ago—it was the object of envy. But in the last generation its character changed. The penthouses at the top keep getting larger and larger. The apartments in the middle are feeling more and more squeezed, and the basement has flooded. To round it off, the elevator is no longer working. That broken elevator is what gets people down the most” (quoted by Luce 2010).

The policy challenge in the advanced countries is that sustaining current levels of openness will require addressing these domestic challenges at the very time when growth could be slowing and the ability to effect redistribution is being impeded by broader medium-term fiscal concerns. In this light, the changing attitudes to globalization and free trade cited above are not surprising.

We focus here on what is now different in the West’s ability to sustain globalization. A starting point is the view, described in Rodrik (1998), that sustaining openness requires a domestic social consensus in its favor, which in turn requires mechanisms of social insurance to cushion domestic actors against globalization-induced shocks. Rodrik (1998) provides evidence showing that this domestic consensus can be captured in the relationship between the size of government (a proxy for social insurance mechanisms) and openness.

More direct evidence of the importance of social insurance comes from a paper by Autor, Dorn, and Hanson (2013), who show that rising exposure to Chinese imports increases unemployment, lowers labor force participation, and reduces wages in local labor markets. They estimate that the exogenous component of this shock explains one-quarter of the contemporaneous aggregate decline in US manufacturing employment. They estimate that rising exposure to Chinese import competition explains about 16 percent of the decline in US manufacturing employment between 1991 and 2000 and 27 percent of the decline between 2000 and 2007. Transfer payments for unemployment, disability, retirement, and health care also rise sharply in exposed labor markets. They estimate the increase in annual per capita transfers attributable to rising Chinese import competition at $32 in the first 10 years and $51 in the last seven years of the sample, which translates into total expenditure of about $5 billion in the 1990s and almost $15 billion in the 2000s. The deadweight loss of financing these transfers is one-third to two-thirds as large as US gains from trade with China.

Can the West sustain these social insurance mechanisms? According to Summers (2008a), globalization both increases the need for social insurance and undermines the government’s ability to provide it, because it renders more factors, especially capital and high-skilled labor, more mobile and less easy to tax.
Has capital become less easy to tax? Figure 4.1 plots the marginal effective tax rates on capital in some important OECD countries and for the OECD as a whole. These rates have been sharply declining, and there is little pressure to reverse these trends.

For the OECD as a whole, the average marginal tax rate declined from about 55 percent to almost 40 percent, a 15 percentage point decline. These declines were witnessed across most, if not all countries. In the United States, rates declined from 65 percent to just over 50 percent; in Germany they fell from about 60 percent to less than 50 percent. Of course, these declines reflect pressures other than globalization and the attendant difficulty of heavily taxing mobile capital, but these pressures have been important.

A new development adds to the problems. Across the OECD, the share of the economic pie accruing to capital has been increasing from about 35 percent to 40 percent in the last few years (figure 4.2). This increasing share has prompted several commentators, including Krugman, to argue that the debate about inequality and trade and inequality in the 1990s, which related to inequality within types of labor (skilled versus unskilled), should now be viewed through a different lens, because inequality is increasingly between capital (and those who own it) and labor.

For the purposes of our argument, what is important is this: Not only is the ability to finance mechanisms of social insurance being undermined by weak growth and the burden of debt (Ruggie 1998); slippery, mobile capital is now accounting for a larger share of the economic pie. The funding of social insurance through taxation is thus going to become more difficult.

The Good News: The Protectionist Dog That Barked but Did Not Bite

Several commentators have remarked on the fact that despite suffering perhaps the biggest global trade shock in the recent global financial crisis, the world did not succumb to protectionism. This response stood in stark contrast to the experience of the 1930s. Explanations for the difference have included the facts that (a) countries could use macroeconomic policy instruments (monetary and exchange rate), which adherence to the gold standard initially prevented in the 1930s (Eichengreen and Irwin 2009); (b) automatic stabilizers were in place, by way of transfers and unemployment benefits (Autor, Dorn, and Hanson 2013); and (c) the deeper integration created by modern production chains rendered protectionism self-defeating (Baldwin and Evenett 2009).

The bigger puzzle is this: How did the West, and the United States in particular, adjust to arguably the biggest structural trade shock in its history—namely, rising imports from China—without any serious recourse to protectionism? Why was there less protectionist outrage in the United States against China than there was against Mexico in the 1990s or Japan in the 1980s? The domestic uproar against China did not match the backlash created in the context of the North American Free Trade Agreement (NAFTA),
and actual protectionist actions did not come remotely close to the actions taken against Japan (the Reagan era witnessed the greatest upsurge in trade barriers in the postwar period; see Destler 1992).

The differences cannot be explained by the relative magnitude of the three shocks, as the Chinese shock was orders of magnitude larger than the early shocks. Figure 4.3 plots imports from Mexico, Japan, and China as a share of US domestic consumption between 1962 and 2011. At their peaks, Japan accounted for 3.6 percent of U.S. consumption, whereas China accounts for about 5.2 percent.  

Table 4.2 quantifies the trade shocks to the United States from the three countries. The shock is computed in three ways (each scaled by the working-age population in the United States or the domestic consumption of manufacturing): (a) average imports over the relevant period (for convenience, all shocks are considered to extend over a 20-year period: Japan 1970–90, Mexico 1980–2000, and China 1990–2010); (b) the change in imports over the period; and (c) both average changes and changes calibrated by per capita GDP in each country.

As table 4.2 shows, depending on the measure used, the Chinese shock was either 4 to 5 or 10 times as great as the Japanese and Mexican shocks. Calibrated by per capita GDP, it was even greater. (One reason to calibrate by per capita GDP is that trade with low-income countries is of the Heckscher-Ohlin variety. It therefore imposes greater domestic political costs [than, say, trade in similar goods between countries at similar levels of development], in particular because these costs are disproportionately borne by unskilled labor, which competes more directly with foreign imports.)

Several explanations are possible for the differential response to the China shock. One could be that the measure of recorded imports exaggerates the trade shock because of the difference between gross flows and value-added flows. Chinese exports embody less value added than the exports of many other countries because of the large volume of intermediate inputs it imports and transforms into exports. Even making allowances for this distinction, however—and the problem was arguably as acute in relation to Mexican maquiladora exports to the United States—would hardly make a dent in the numbers presented above.

A second explanation could be that in the case of Mexico, the uproar was exaggerated because there was a focal point: a trade agreement that had to be passed by the US Congress. But in the case of China,

---

28. Appendix figure A.3 plots the same data but for a shorter period for which value-added trade data can be computed. Gross exports overstate value-added exports for China, but they overstate them even more for Mexico.
29. Trefler (1993) shows that cross-industry differences in protection are associated with the change in import penetration, not its absolute value. Autor, Dorn, and Hanson (2013) use import penetration as a share of working-age population as the measure of trade shock.
31. In appendix table A.1, the figures for China are recomputed based on value-added trade data (we cannot do the same for the Mexican and Japanese shocks, which would bias the comparison in favor of understating the China shock). The size of the Chinese shock declines, but it remains orders of magnitude larger than the earlier shocks from Japan and Mexico.
there has also been an identifiable target and identifiable policies: currency manipulation. Moreover, Mexico was an ally, whereas China is a potential adversary and competitor to big power status, which should have increased the outcry and concerns in domestic US politics.

A third explanation is that by the time the China trade shock arrived, the United States had specialized so much away from unskilled labor that there was less to disturb domestically. For example, the number of workers employed in the US clothing sector declined from 900,000 in 1990 to 150,000 in 2013. In technical terms, the United States is no longer in the cone of diversification (Edwards and Lawrence 2013). The estimates of employment disruption by Autor, Dorn, and Hanson (2013) for the Chinese case suggest that this argument cannot be a full, or even an important, explanation, however.

A fourth argument for the relatively muted domestic response is that the size of the Chinese market and the strategy of openness to US FDI essentially co-opted US companies and capital, which had an incentive to support rather than criticize China. The Japanese experience was different from the Chinese experience in two important respects: Trade conflict with Japan reflected head-to-head competition in some specific industries (steel, cars, semiconductors) rather than conflict based on unequal endowments. It was US capital rather than labor that was the victim in the Japan episode; as it had unusual influence in the political process, there was correspondingly more of a response. In addition, Japan had not created the same stake for US companies in Japan as China had.

Finally, it is possible that the underlying macroeconomic situation was better when the Chinese export juggernaut arrived.

One conclusion from all this is that if US domestic politics could survive a shock as great as that from China, there may be an underlying resilience (helped considerably by government insurance mechanisms) that should not be underestimated. Moreover, it could be argued that structural shocks similar to China’s are unlikely to repeat themselves. This fact should temper unremitting pessimism about the future of globalization.

One can generalize the Chinese experience in the United States more broadly to other advanced countries. Figures 4.4 and 4.5 illustrate the change in OECD country imports in favor of developing countries. They plot the average income level of manufactured imports into the United States, Japan, and the Europe Union. The per capita GDP level of each source country, measured relative to that of the importing country, is weighted by its share in total manufactured imports of the reporting country (in figure 4.4, the per capita GDP and import share are contemporaneous; in figure 4.5, the per capita GDP is fixed at the 1980 level). In all cases, imports from the early 1990s are being sourced progressively from poorer countries, suggesting an increase in competition from lower-wage countries. In the European Union, for example, the average income level of imports drops from 100 percent to 75 percent. The
point is that all advanced economies have experienced large trade shocks, without recourse to serious protectionism.\textsuperscript{32}

\textbf{The China Challenge: Bretton Woods Rules or Hyperglobalization Rules?}

China will play a critical role in shaping the future of globalization, just as the United States did in the immediate aftermath of World War II. Its economy is as large as the United States’ (in purchasing power parity terms), and its merchandise trade is larger. Over time, unless China implodes, the differential in economic strength will widen in its favor. Under reasonable assumptions about growth, China will become the dominant economic trader, accounting for twice as much trade as the United States and four times as much as Germany in 2030.

The China challenge will be twofold. As it becomes the world’s largest economy and trader, its markets will become more important for other countries, especially low-income ones. China will thus matter to globalization not just as an exporter, but also as an importer.

This tension has been best exemplified in recent years by China’s mercantilist growth strategy. In this millennium, China’s growth and exports have been underpinned by an undervalued exchange rate, a policy that has been successful for China from a domestic perspective. By increasing the global export supply and depressing the global price of tradables, this policy may have set back the diversification possibilities of other poorer countries, however (Mattoo, Mishra, and Subramanian 2012).\textsuperscript{35}

The China challenge is more broadly applicable to middle-income countries. On the one hand, these countries will continue to rely on trade and foreign markets as a means of increasing their growth. On the other hand, these countries and their markets are becoming big enough (as the projection for China suggested) to offer opportunities for the average citizen in other poorer parts of the world. If China’s actions are market opening, there will be little conflict between its domestic imperatives and the demands of an open system. If, however, it uses its policy space to implement beggar-thy-neighbor policies, conflict with other countries, including poorer ones, will ensue.

A second, possibly more important challenge from China’s rise will pertain to the openness of the global system. After World War II, the United States initially bequeathed an open, rules-based trading system. Subsequently, reasonably successful cooperation between the two dominant trading powers—the United States and Europe—was achieved in the GATT/WTO through reciprocal exchange of market-opening commitments. Can this same mechanism be effective going forward with China as a dominant

\textsuperscript{32} The value-added counterpart of figures 4.4 and 4.5 are appendix figures A.2 and A.3. The broad trends remain the same.

\textsuperscript{33} Undervalued currencies are in effect both an import tax and an export subsidy; countries that maintain them wind up reducing the profitability of industries in countries with which they trade.
trader? Will it be possible to get China to reduce its policy barriers, especially in areas not covered by WTO rules, such as government procurement, investment rules, technology indigenization, and services sector policies?

There may indeed be a structural problem limiting the scope for reciprocity. When the United States and European Union dealt with each other in the postwar period, their markets were largely open; in areas where they were closed, they were broadly closed to the same extent. This symmetry facilitated reciprocity. In the future, problems could arise because the United States and European Union are, in policy terms, more open than China (see figure 2.10 above).

China is highly open in terms of trade outcomes, and it has made great strides in removing policy barriers as part of WTO accession. It is, however, more closed in policy terms, especially outside the traditional goods area. In services, technology, and government procurement, closed policies take the form of continuing state control over a large share of economic activity. According to Borchert et al. (2012), China’s services sector policies are about three times as restrictive as those of the United States. For example, in the retail sector, China limits foreign ownership to 49 percent if the retailer has set up 30 or more stores that sell multiple types and brands of goods. In any future bargain, the United States will, by virtue of previous liberalization, have less to offer China.

The paradox will be that China will have greater leverage in bargaining by virtue of its much larger volume of trade but will also have higher trade policy barriers. It is as if in a duel, one party offers a smaller target and has a pistol with longer range. By 2020, China’s imports will be 1.5 times larger than the United States’ imports and twice as large as Germany’s, conferring the kind of power that comes with being able to determine access to its markets. Future bargaining will therefore be structurally imbalanced in China’s favor, making reciprocity more difficult to achieve.

This structural imbalance will be a persistent source of tension between the United States and China. As US manufacturing sector hollows out, and the United States comes to rely to an even greater degree on services, it will seek to open markets overseas, especially in China. If China’s future opening is slow, over time the United States may be increasingly tempted to play the unfairness card based on the disparate levels of policy openness: Why should our markets be more open than those of a rival and equal?

This imbalance in bargaining could remain at the level of sparring and skirmishing without systemic consequences. But suppose failures to address its structural economic problems—stagnating median household income, rising inequality, declining economic mobility—creates a large, disaffected, and beleaguered middle class in the United States and that the intellectual consensus in favor of openness becomes increasingly frayed, as it has in the last few years. Frustrated by China’s unwillingness to open the new sectors of its economy and lacking the carrots to overcome its unwillingness, spurred by a weak economic climate and shifting intellectual certitudes, and goaded by perceptions that China is not making
its fair contribution to keeping markets open, the United States could be tempted to threaten to close its own market to China unless China further opens its own. In this scenario, especially if China cannot wean itself off mercantilism and state capitalism, the scope for trade conflict and tension could increase considerably, jeopardizing the openness of the global system.

Concerns about China’s trade policies have not been confined to rich countries. The Chinese export juggernaut is a source of concern across the developing world. Brazil, India, Mexico, Korea, South Africa, and other emerging market countries chafe under China’s mercantilist exchange rate policies. Most of the antidumping actions taken by developing countries have been against Chinese imports. And one of the dirty secrets of the Doha Round is that its collapse was caused in part by the reluctance of emerging market countries to liberalize their economies and expose themselves to Chinese competition.34

So a big, first-order question for the open system is how to prompt China to adopt more open policies. The problem becomes acute if one recognizes China’s economic dominance and the degree to which other countries have lost leverage to influence Chinese policies. A little over a decade ago, the West could essentially determine the terms of China’s accession to the WTO; that is no longer the case today.

The Changing Global Governance of Trade: Mega-regionalism Meets China

For a variety of reasons, regional integration has been, along with unilateral liberalization, the preferred mode of liberalization. As a result, the governance of world trade has shifted decisively toward regionalism and away from multilateralism.

So far, preferential trade agreements (PTAs) have been North-South “deep” regional trade agreements, occasioned in part by the rise of supply chains. A Doha Round that was midwifed in an unusual bout of post-9/11 world solidarity and saddled with an agenda of issues like agriculture and tariffs that have less relevance now than new issues (exchange rate mercantilism, services, government procurement, investment, export restrictions) also contributed to a preference for regionalism (Mattoo and Subramanian 2012a and Baldwin 2011b).

North-South PTAs have not posed a serious threat for globalization; in fact, they may have contributed to significant opening through a process of competitive liberalization. But this relatively

34. In 2009, China’s share of imports in the most protected sectors was substantially larger than its share of overall imports and dwarfed that of any other supplier. Its share in these sectors was more than 70 percent in Japan; more than 60 percent in Korea; about 55 percent in Brazil; and about 50 percent each in the United States, Canada, and the European Union. Even in these protected sectors, China’s share increased dramatically over the course of the Doha Round. In many importing countries (for example, Brazil, the European Union, and the United States), China’s share more than doubled. Also striking is how much market share China has gained even in countries such as Canada, Mexico, and Turkey, which have free trade agreements with close and large neighbors. Thus, liberalization under the Doha agenda, especially in the politically charged, high-tariff sectors, is increasingly about other countries opening their markets to Chinese exports (Mattoo and Subramanian 2012a).
benign outcome cannot be taken for granted in the case of the mega-regionals looming on the horizon. On the one hand, they will involve deep integration and facilitate further globalization. After all, the world trading system is already multilayered in terms of levels of integration (the European Union, the European Free Trade Agreement, customs unions, other free trade agreements, and so forth). A Trans-Pacific Partnership and Transatlantic Trade and Investment Partnership will just add another layer of integration.

On the other hand, such agreements could be exclusionary, depending on the extent to which they conform to the principles of open regionalism. Their effect will depend, above all, on how outsiders—the middle-income countries, especially China—react. If China views these agreements as economic war and containment by other means, and retaliates by concluding its own regional agreements, excluding the large traders, fragmentation and conflict could lie ahead.

It is not that the WTO and multilateralism have become totally irrelevant. More and more countries, including Russia, want to join the WTO. Its dispute settlement system functions effectively, its basic rules are broadly respected. The question is whether it retains its relevance as a key forum for facilitating further liberalization or transforms itself into an institution that serves mainly as a court of trade law and an overseer of regional trade.

The (Non) Challenge of Low-Income Countries

The antiglobalization and anti-WTO crusade of the 1990s (culminating in the protests in Seattle in 1998) forced policymakers into a strenuous defense of the development friendliness of the trading system. The Doha Round, for example, was formally dubbed the Doha Development Agenda, which seems an overdone title.

The perception that the trading system is unfair to low-income countries stems from the fact that rich country trade barriers are highest in agriculture and low-skilled manufacturing (textiles, clothing, and footwear), which tend to be important exports for low-income countries. This situation changed over the last decade or so. Rich country barriers in these sectors declined in the aftermath of the Uruguay Round, with Canada, the European Union, the United States, and Japan all significantly improving preferential access to low-income countries, in terms of both country and product coverage (the European Union’s Everything But Arms and the United States’ Africa Growth and Opportunity Act are two notable examples). In some cases, these schemes also became less arbitrary.

As a consequence, low-income countries are, at worst, treated no worse than the typical country and, at best, treated better. Nearly all low-income countries (“least developed countries,” in the jargon of the trading system) face lower trade barriers than most other countries, because they receive generous preferences (albeit with some problems, depending on which country is granting them). There are some exceptions to these preferences (Bangladesh and Cambodia in clothing; West African countries facing
rich country cotton subsidization). But even these exceptions create a situation only of parity with other countries. Increasingly, larger developing countries have also started granting preferential access to low-income countries, although the coverage and magnitude of this access are limited.

Hoekman and Nicita (2011) calculate that the average barriers facing exports of sub-Saharan Africa are very low and much lower than for other countries. The average level of restrictiveness (including nontariff measures) that exports from sub-Saharan Africa face in other markets—4.4 percent in high-income countries and 6.0 percent in upper-middle-income countries, taking account of preference margins—is consistently lower than for any group. The comparable numbers are 6.3 and 15.6 percent for high-income countries as exporters, and 5.7 percent and 11.8 percent for upper-middle-income countries as exporters. (An exception is South Asia, which faces higher barriers because its exports face higher MFN tariffs and because countries such as India and Bangladesh are not included in the major preferential schemes, especially in the United States.)

Low-income countries also receive considerable space to pursue their own policies. In relation to rules on subsidies, intellectual property, and local content requirements, the least developed countries face weaker obligations. Moreover, the thresholds for taking contingent protectionist action (countervailing and antidumping) against exports of low-income countries are generally higher.

So it is not clear what more could be done for low-income countries internationally that would materially alter their growth prospects (box 4.1). Their growth challenges are predominantly domestic, as indicated in Rodrik (2013). Even recognizing the important point that Baldwin (2012) makes—that these countries need to get on to the new supply chains—it is not clear what other countries might do to galvanize this process, especially if the supply chain phenomenon is to some extent about geography.

For low-income countries, a trading system that allows them policy space to pursue appropriate growth strategies and that at the same time keeps global markets open for their exports is critical. After all, such was the external environment that allowed today’s middle-income countries to prosper. Despite the tightening of trade rules since the formation of the WTO, there remains enough policy space for these rules not to become a straitjacket for today’s low-income countries. Apart from some specific issues (such as food security), the real concern will be whether the external environment will remain as benign as in the previous two decades, so that low-income countries can export their way to growth. The actions of high- and middle-income countries will be critical.

5. POLICY RESPONSES

The key challenges facing globalization are sustaining domestic support for it in the West and ensuring that China continues to open its markets. Open markets in China are an important part of China’s domestic agenda for sustaining convergence. They also offer opportunities for poorer countries similar to
those that China enjoyed in industrial countries over the last few decades. They are critical to keeping the trading system open and free of serious conflict. The challenge of mega-regionalism is related to the China challenge.

**National Responses**

Sustaining and furthering globalization will be determined at the national rather than the international level. For the United States and Europe, actions are needed to revive growth and address fiscal challenges, especially the challenges stemming from growing entitlements. For the United States, there is
the additional challenge of addressing the problems of stagnating wages, rising inequality, and declining mobility. Success on these fronts will provide a surer guarantee that globalization will proceed apace. Several studies note that the demand for trade protection is inversely and robustly related to the state of the economy (see Rodrik 1998).

For its part, China should have a stake in preserving the open system for the simple reason that its rapid economic transformation over the last three decades was predicated crucially on openness. That transformation is still far from complete: China’s standard of living is still only 20 to 25 percent that of industrial countries. Completing that transformation is critical for the political legitimacy of China’s policymakers. In these circumstances, disrupting the open system would amount to biting the hand that has fed China and its rulers.

Indeed, going forward, the Chinese agenda for reforms, as elaborated in the government-imprimaturmed 2012 World Bank report *China 2030*, should be entirely consistent with an open system: China’s domestic needs are broadly outsiders’ wants. For example, the nontransparent practices of the state enterprises and the financial repression and closed nature of China’s capital account are a big concern for foreign firms trying to access the Chinese market and for firms around the world trying to compete with an undervalued Chinese currency. The more China reforms its state enterprises and state-owned banks, the easier it will be for foreigners to do business with China.

*China 2030* calls on China to move more toward an innovation-based economy, which would require stronger protection of property rights, another key demand of outsiders. China needs to reduce its pollution and move toward a more carbon-efficient economy, which would allow it to play a constructive role in global climate change efforts. In all these cases, tensions will undoubtedly arise from differing senses of urgency about specific actions. But across the board, there is no fundamental conflict between what China needs to do domestically and what it needs to do to sustain an open system.

Of course, China could falter because its domestic problems—rising inequality and corruption, increased demands for accountability and participation, environmental deterioration—cannot be easily resolved. In this case, globalization would suffer.

**International Responses to the Challenges Facing the West**

International/collective responses can help in relation to both these challenges. In relation to the first, the increasing mobility of capital and its ability to escape taxation needs to be addressed. Rodrik (1997, 81) argues that this phenomenon “undercuts the revenue sources needed to maintain social and political cohesion and ultimately erodes support for free trade.” Two new developments have exacerbated this problem: Capital has become more mobile (reflected in growing financial globalization and increased FDI flows), and the distribution of income in most OECD countries has moved substantially in favor of capital (and also in favor of highly skilled people), increasing the size of the tax base that can elude taxation.
In many emerging markets as well, especially China and India, capital is accounting for a larger share of the pie (figure 5.1). The global tax base (not just that of the OECD) is becoming more slippery. If countries and companies exploit the mobility of capital, the global ability to provide social insurance will decline, creating problems for globalization. Hence, there needs to be much greater cooperation between rich and emerging market countries (and, of course, tax havens) on how to tax capital and how to share the taxes from capital (Summers 2008b). This cooperation can take the form of greater harmonization (which would be difficult and entail a degree of regulatory convergence that countries will find difficult). Or it can take the form of countries doing their best to allow other countries to better enforce their own tax rules (a recent example involved Switzerland relaxing its secrecy laws to allow the United States to go after its tax evaders).

**International Responses to the Challenge of China and Mega-Regionalism**

The China challenge is a broader problem of cooperation in the face of a shift in economic power from the United States and Europe toward a rising one. It is in this context that even the new attempt at mega-regional agreements (Trans-Pacific Partnership and Transatlantic Trade and Investment Partnership) must be seen. At one level, these agreements are about deepening integration at a time when the multilateral liberalization process has become moribund and the Doha Round remains in cold storage. At another level, the issue is how the United States and European Union deal with the rise of China.35

At the risk of overgeneralizing, the challenge in the trade arena can be summarized as follows: China is happy with the status quo and the United States is not. China—and the other larger emerging markets, such as Brazil, India, and Russia—is reasonably content to have Bretton Woods rules apply to it and hyperglobalization rules apply to its large partners. China will liberalize and open up its markets in line with domestic rather than external imperatives. Its partners, especially the United States, will increasingly refuse to acquiesce in this status quo. But given China’s dominance and the weakness of the United States, the United States’ ability to force or induce China to change will be limited. The mega-regionalism demarche of the United States is an attempt to exert pressure on China.

How can these differing perspectives and positions be reconciled? The larger partners of the United States and China need to deploy a strategy that takes account of the possibility that China may occasionally be tempted into a less than-benign economic hegemony while reinforcing its incentives to act to preserve an open economic system.

The possibility of the misuse of hegemony would not be unique to China. It was famously said of the United Kingdom that Britannia ruled the waves by waiving the rules. The United States also

---

35. In the case of the Trans-Pacific Partnership, containment is more political. With the Transatlantic Trade and Investment Partnership, agreement by the United States and the European Union to common regulatory standards would preempt China from imposing its standards in international markets and even force China to adhere to these common standards.
occasionally succumbed to this temptation. In 1955, it excluded agriculture from GATT disciplines. In the early 1970s, it unilaterally blew up the Bretton Woods system when it became a straitjacket on domestic US policies. In the 1980s and 1990s, it cajoled and coerced developing countries to take on costly obligations (relating to intellectual property and capital flows, for example). In the dark Nietzschean view that “power is never held in innocence,” misuse of hegemony is intrinsic to hegemony. It is inevitable given the infinite capacity for countries to succumb to the delusion that John Adams memorably warned about, that “power always thinks it has a great soul.”

History suggests that the best defense against hegemony is multilateralism, which offers a modicum of protection for the weak against the dominant power. Keeping China tethered to the multilateral system, in which the United States and other major countries can exercise some countervailing influence, offers the best insurance against its unrestrained exercise of hegemony.

Multilateralism could work as a defense against China in several ways: in shaping rules, in promoting adherence to them, and more broadly in defining legitimate behavior. With China’s growing size, the balance of negotiating power will be with China rather than its partners. Multilateralism ensures that China’s trading partners will have enough heft to negotiate with China in a more balanced manner. For example, China might be willing to open its markets in return for the United States, European Union, India, and Brazil opening theirs. Its willingness to open up in a similar manner in negotiations with just the United States, or European Union, or with some less powerful combination is far from clear.

A similar argument carries over to enforcement and the incentives to adhere to previously agreed upon rules. China’s incentive to abide by multilateral rules will be stronger than its incentive to abide by a series of bilateral agreements, because the reputational costs of being seen as errant are much greater in the multilateral context. The opprobrium of being a deviant from multilateral norms is China’s great fear, rendering multilateralism the best weapon the world can deploy against a dominant China.

These arguments for multilateralism have an important corollary for the United States and other countries. They imply less recourse to bilateral and regional dealings with China and with each other. The more countries elevate the role of bilateralism in dealings with China, the less China will be anchored in the multilateral system and the more exposed countries will be to the exercise of Chinese dominance. One operational consequence, advocated by Lieberthal and Wang (2012), is to expand the US-China strategic economic dialogue to include the larger countries in the world—Europe, Japan, Brazil, and India, for a start—whose heft can be an effective counterweight to that of China.

These arguments in favor of multilateralism and against regionalism and bilateralism apply across the board in the fields of currency, finance, and trade. But they carry particularly important implications in the field of trade because of the current environment, in which the WTO appears moribund and regional initiatives are flourishing. The recent trade initiatives of the Obama administration—the Trans-Pacific and Transatlantic Partnerships—are regional initiatives. As box 5.1 suggests, they are particularly fraught in the context of a rising China.
In the old debate between regionalists and multilateralists, the divide was not about the end point: All parties wanted global free trade. Rather, the divide was over whether regional agreements would be, in Jagdish Bhagwati’s words, a building block or a stumbling block in the way of that goal, with the regionalists falling in the former category and the multilateralists in the latter.

Regionalists would point to the evident success of regionalism in achieving deeper liberalization: In many or most cases, border barriers have been eliminated in goods and services, and in some, behind-the-border barriers have also been addressed. But both forms of regionalism involving China render this old debate less relevant.

No major country has yet embraced regional agreements with China (although the Association for Southeast Asian Nations [ASEAN] has embarked on this path). The arguments against regionalism are really the flip side of the arguments for multilateralism: negotiating with China will lead to agreements that are weighted in favor of China, because it has bargaining power. If the basic problem is the imbalance of leverage arising from China’s size, regionalism will by definition be less effective than multilateralism. For the same reason, getting China to adhere to these agreements will also be difficult.

But countries are increasingly negotiating agreements around China, with the Trans-Pacific Partnership representing the best recent example. Advocates of regionalism have long relied on the competitive dynamic it creates: If two countries negotiate preferential reductions of barriers, one or several outsiders will be hurt. These outsiders will then have an incentive to negotiate preferential agreements themselves. This process will continue until the goal of global free trade is achieved—or so went the Bergsten-Zoellick theory of regionalism as promoting competitive liberalization.

But consider three ways in which the Trans-Pacific Partnership might play out. In the first, the United States embarks on a process of deep integration with a number of Asia Pacific countries without China. To avoid the dangers of hostile regionalism (that is, excluding China), Trans-Pacific Partnership countries could subscribe to the principle of open regionalism: Countries get all of the agreement’s benefits only if they embrace its terms.

The problem with this approach is that China would never agree to fall in line with rules it did not participate in negotiating. For example, if Trans-Pacific Partnership members negotiated rules against undervalued exchange rates, China would probably stay away. If it did, the agreement would hardly achieve the objective of disciplining problematic Chinese policies that adversely affect the open character of the trading system.

In the second scenario, the United States invites China to the Trans-Pacific Partnership negotiating table to be part of the process of creating the rules. Would its participation really be superior to negotiating with China multilaterally, where the European Union, Brazil, and India would also be at the table? If the problem of a rising China is that it will have enormous bargaining power by virtue of its economic size and dominance, then a multilateral process will add more negotiating heft on the other side of the negotiation. How can it not help to have Brazil, India, and Europe as part of the group putting pressure on China to create better rules and to adhere to them? Of course, there is no guarantee that Brazil and India will always be on the side applying pressure on China. But where important interests are at stake, they are likely to do so.

In the much worse third scenario, China construes the Trans-Pacific Partnership as an act of hostile regionalism (Yao Yang’s [2013] reaction is typical of this perception) and negotiates preferential agreements of its own—with, say, the European Union alone. Such a situation would create significant trade diversion for the United States and other exporters because of high Chinese levels of protection in certain areas. The Trans-Pacific Partnership could thus provoke China into playing the regionalism game in a way that could fundamentally fragment the trading system. Down this path lies the folly of the interwar years.

(continued on next page)
Box 5.1 Multilateralism versus regionalism: The risks from the Trans-Pacific Partnership (continued)

In sum, the Trans-Pacific Partnership will either exclude China (and open regionalism may have little sway in persuading China to join) or be less effective in engaging China, because it would exclude other large trading countries (the European Union, Brazil, and India), whose collective heft might be crucial in balancing the bargaining power of China. The lesson is that the success of regionalism in reducing barriers and generating the competitive dynamic for further liberalization simply cannot be applied to China. The successes of regionalism typically involved a big economic power (the United States, the European Union, Japan) negotiating with smaller countries. The smaller countries did most of the incremental liberalization, because the larger countries held the balance of negotiating power and influence. With China, the power balance is reversed, rendering many of the old arguments for regionalism obsolete.

Regional and discriminatory solutions carry greater risks. The challenge of anchoring China in the multilateral trading system—as well as providing a fillip to growth in industrial countries through further liberalization—can be addressed by embarking on a new and comprehensive multilateral initiative. This initiative would anticipate the changing interests and concerns of all the big trading nations in a way that the Doha agenda did not. It would also pave the way for a reciprocal liberalization mechanism—you open your markets in return for my opening mine—that has been the basis for previous successes in the trading system.

To achieve this reciprocity, first and foremost, the world should declare that the Doha Round is dead and place a wider range of issues on the agenda. China’s trading partners remain concerned by Beijing’s exchange rate policies, as well as the protection and discrimination that stem from China’s state capitalism. China and other countries have an interest in ensuring that their exports are not subject to antidumping and trade restrictions, uncertainty from investment regulations, and international rules on subsidies in relation to climate change.

Everyone has an interest in preventing export protectionism, liberalizing trade in goods and services, and opening government procurement markets. To achieve these goals, Mattoo and Subramanian (2012a) call for a new China Round of multilateral negotiations focused on some of these issues, with participation (initially) by a core group or critical mass of large trading countries (Low 2012).

Any new initiative will have to break from the past in one key respect. Countries in the West have been the drivers of past trade negotiations. China and the other big emerging market countries must now take the lead in negotiating further multilateral liberalization (the alternative, an orderly retreat from globalization, is explored with some skepticism in box 5.2). If they do not, there is the risk that mega-regional agreements will spread, which would be detrimental to the excluded larger emerging markets.
Box 5.2  Is an orderly retreat from globalization possible?

In discussing the challenges of globalization, the possibility of a retreat from it must not be ignored. A meaningful and consequential retreat would have to be led by the major trading powers, especially the United States, Europe, and China. If it is sparked by political or security conflicts, there can by definition be no orderly retreat.

One recent historical example did represent a retreat from globalization. Although not close to the severity of protectionism seen during the 1920s and 1930s, the 1980s experience of US-Japan trade relations is nevertheless instructive. In the wake of the recessions in the early 1980s and the appreciation of the dollar through 1985, the United States became more protectionist (Destler 1992). This protectionism took several forms. Some actions were consistent with the letter if not the spirit of international rules (for example, recourse to antidumping and countervailing duties). Other actions clearly violated the rules (for example, getting Japan to impose voluntary export restraints) or involved the threat of illegal trade sanctions to persuade/coerce Japan to open up its own market. This experience suggests that it will be difficult to craft rules ex ante that bind the very players that have the power to violate or disregard them in the very circumstances that create the greatest incentives to do so.

6. CONCLUDING THOUGHTS ON THE FUTURE OF TRADE INTEGRATION AND COOPERATION

Can and will the ongoing process of hyperglobalization of goods and services continue? Or, to use Keynes’ evocative phrase, might there be “serpents to the paradise” of the ongoing process of hyperglobalization?

Why Optimism About Globalization?

Although trade has been rising rapidly, the process is less than half complete. On a value-added basis, the world trade-to-GDP ratio is about 25 percent, but a simple frictionless gravity model predicts that the theoretical maximum should be substantially greater. At least three forces will drive globalization toward and sustain it at higher levels: economic convergence; technology; and interests, ideas, and institutions.

Economic Convergence

As more countries continue to grow and to grow more rapidly, trade will increase, as figure 2.4 illustrates. Clearly, the pace of globalization will be affected by the pace of convergence. If Rodrik’s more sober assessment prevails, the pace of globalization may slow, but it will not be reversed. If the pace described in table 2.1 is sustained, ongoing hyperglobalization will be as well.

Technology

Predicting the pace of technological progress is impossible. Revolutions in transportation, and then in information and communication technologies, have driven trade globalization. Even if the pace of new

36. Recall that this model suggests that the ratio of world trade-to-GDP should be 1 minus the sum of squared shares of countries in world output. With convergence and a sufficiently large number of countries, the sum of squared shares should converge to zero and the ratio of world trade to GDP should converge to one.
discoveries slows, there is scope for the spread of existing technologies, both directly and embodied in FDI. Mobile telephony, internet usage, and connectivity are still far from universal (Aker and Mbiti 2010).

**Interests, Ideas, and Institutions**

Bhagwati (1988) identifies a set of factors he calls the three Is: interests, ideas, and institutions. The very fact of hyperglobalization deepens the enmeshing of interests across countries, people, and companies. In the current phase, the additional reinforcing factor relates to the phenomenon of criss-crossing globalization discussed earlier.

One of the widely noted features of the global financial crisis was the drop in trade that exceeded the decline in the aftermath of the Great Depression. Yet unlike in the past, the world did not collapse into a protectionist spiral. One reason why this collapse did not occur was that countries, no longer tied to the gold standard or otherwise straitjacketed, had broader macroeconomic policy options. Another reason was that the vertical integration of production via supply chains made it difficult and undesirable for countries to impose barriers that would undermine these chains and hence trading opportunities.

The supply chain phenomenon has a broader counterpart. Not only are goods moving back and forth—capital flows are, too. FDI flows no longer just flow downhill from rich to poor countries. Brazil, China, and India are all becoming large exporters of FDI. As capital relocates internationally, the political economy of protectionism also changes. US firms in China have lobbied strongly against US trade action against China. India can now less afford to repel FDI if Indian companies at the same time seek to operate in and from foreign markets (Mattoo and Subramanian 2010).

The recent crises provoked an existential debate about capitalism and finance, but the ideological near-consensus that trade in goods and services as well as FDI should flow relatively unimpeded has not been dented.

**Is There Reason to Be Sanguine About Trade?**

The cardinal sin of forecasting is to extrapolate the recent past, as Norman Angell, future Nobel Peace Prize winner, did in 1910, when he published *The Great Illusion*. This pamphlet-turned-book acquired cult status for propagating the view that Europe had become so interlaced economically through trade, credit, and finance that war was impossible. Twentieth century wars would be so economically devastating even to the aggressor that waging one would amount to self-inflicted folly. 

37. Global Trade Alert (2012) suggests that there was an upsurge in protectionism after the global financial crises, the quantitative impact of which remains far from clear. Hufbauer et al. (2013) document the rise of local content requirements in a number of countries.

38. In the words of Lord Esher, Angell’s most earnest disciple, the inevitable consequences of “commercial disaster, financial ruin, and individual suffering” would be “pregnant with restraining influences.”
Notwithstanding the five influences discussed above, history’s lesson is that we cannot be 100 percent certain that the enmeshing of interests will be strong enough to sustain the status quo. Nor is there a cast-iron guarantee that the current ideological embrace of markets as the predominant basis for organizing economic relations will survive the vicissitudes of intellectual fashion and the selective and self-serving interpretations of policymakers. There is tail-side risk (that is, a small, but nontrivial probability of catastrophic outcomes) that interests, ideology, and institutions, both domestic and international, will be the inadequate to the task of preserving the current system. And then there is always the unforeseeable and the irrational. World War I, after all, did happen (Subramanian 2011, 170).

Section 5 discussed the factors that become serpents in the paradise of hyperglobalization. They include prolonged weakness in the West, a serious domestic shock in China that precipitates a retreat there, and the vacuum in international governance. The status quo power is in economic decline, and the rising power will prioritize domestic interests over international responsibilities to a greater degree than previous superpowers, because it is still only a middle-income country. Another unforeseeable factor is the politics and projects of militarism and imperialism (for example, a conflict between China and Japan), which could set back globalization.

Most of the actions that will allow positive influences to prevail over globalization-reversing ones will be at the national level: actions to address economic decline in the West and sustain growth in the Rest, especially China. Collective action should help strengthen the institutional underpinnings of globalization. These actions include ensuring that domestic social insurance mechanisms are not undermined by globalization and bolstering multilateral institutions to prevent conflict between the major trading partners. Greater cooperation on taxes may become necessary to preserve funding for these mechanisms. The world should declare the Doha Round dead in order to move to more meaningful multilateral negotiations to address emerging challenges, including any possible threats from new mega-regional agreements. The rising powers, especially China, will have a key role to play to resuscitate multilateralism.

The open, rules-based trading system has delivered immense benefits for all, especially today’s emerging market economies. Preserving it will ensure that low-income countries can also make successful growth transitions. It is often overlooked that the international trading system has witnessed more successful cooperation, especially between the systemically important countries, than the international financial and monetary system. So cooperation to preserve globalization, even if not in its most hyper current incarnation, is of critical importance. It may also prove less difficult.
Table 1.1  Convergence: Growth of developing countries compared to growth in the United States

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>US growth rate of GDP per capita (percent)</td>
<td>1.7</td>
<td>2.47</td>
<td>1.28</td>
<td>0.65</td>
<td>0.02</td>
</tr>
<tr>
<td>World growth rate of GDP per capita (percent)</td>
<td>1.3</td>
<td>2.75</td>
<td>3.17</td>
<td>2.28</td>
<td>1.73</td>
</tr>
<tr>
<td>Number of developing countries in which growth exceeded US rate</td>
<td>2</td>
<td>21</td>
<td>75</td>
<td>80</td>
<td>78</td>
</tr>
<tr>
<td>Percentage of developing countries in which growth exceeded US rate</td>
<td>5.3</td>
<td>29.2</td>
<td>72.8</td>
<td>89.9</td>
<td>83.9</td>
</tr>
<tr>
<td>Average excess over US growth (percentage points)b</td>
<td>0.02</td>
<td>1.53</td>
<td>3.25</td>
<td>2.94</td>
<td>3.03</td>
</tr>
<tr>
<td>Number of countries in sample</td>
<td>38</td>
<td>72</td>
<td>103</td>
<td>89</td>
<td>93</td>
</tr>
</tbody>
</table>

a. Based on GDP in constant dollars. Other columns use GDP in PPP terms
b. Computed as simple average growth of countries whose growth exceeds that of the United States.

Note: Sample excludes oil exporters (as defined by the International Monetary Fund) and countries with populations of less than 1 million.

Figure 2.1  World exports, in current dollars, 1870–2011

Figure 2.2  Stocks and flows of Foreign Direct Investment (FDI), 1913–2011

Sources: Authors, based on data from Bairoch 1996 for 1913–70, Dunning 1983 for stocks and UNCTAD various years for flows for 1970–2011.
Table 2.1  Global tradability of goods and services, 1980–2008 (percent)

<table>
<thead>
<tr>
<th></th>
<th>By shares of world exports (percent)</th>
<th>Tradability (percent)</th>
<th>Evolution of tradability (percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gross trade measure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merchandise</td>
<td>83</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>Services</td>
<td>17</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Value-added measure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merchandise</td>
<td>71</td>
<td>62</td>
<td>57</td>
</tr>
<tr>
<td>Services</td>
<td>29</td>
<td>38</td>
<td>43</td>
</tr>
</tbody>
</table>

Note: We define tradability of a sector as world trade divided by global value added in the sector.
Sources: Authors, based on data from World Bank, various years, and Johnson and Noguera (2012).
Figure 2.3  Dispersion of world output and world exports, 1970–2010

Note: Country equivalents are computed as $\frac{1}{\sum s_i}$ where $s_i$ is the share of each country in world output. A higher number denotes a more equal distribution of output.

Source: UNCTAD, various years.
Figure 2.4  Trade openness, 1870–2010

Note: For 1870–1950, openness is defined using Maddison’s measure of current exports in dollars (deflated by the US consumer price index) and Maddison’s GDP data. For 1951–2010, openness is the variable openk (Penn World Table 7.1) divided by two. Oil exporters and small countries (populations of less than 1 million) are excluded.*

Sources: Maddison (1995); Penn World Table 7.1.

*We chose the openk variable because it is the most comparable with the Maddison (pre-World War II) GDP data in that both are in constant purchasing power parity dollars. For the pre-war export data, there are two options for deflation: a measure of general US inflation (for example, the consumer price index) or a measure of export prices. Maddison provides a real export series based on the latter. We chose the CPI option for the simple reason that the estimates for 1950 (matched better the Penn World Table estimates for the years close to 1950. If we use Maddison’s real export data, the changes over time are even more dramatic than shown in figure 2.4 (i.e., export-to-GDP ratios are lower for the past when exports are deflated by an export price index than a CPI).
Figure 2.5 Two-way Foreign Direct Investment (FDI) flows, 1970–2011

Note: The Grubel-Lloyd index is computed for each country with nonzero positive flows. Each country is then weighted by its share of total FDI flows, either with weights corresponding to the current year (dotted line) or with weights that are fixed at their mean over the period (solid line). The figure shows five-year moving averages (to avoid large spikes).

Source: UNCTAD various years.
### Table 2.2 Merchandise exports as share of world exports by mega-traders, 1870–2030 (percent)

<table>
<thead>
<tr>
<th>Year</th>
<th>United Kingdom</th>
<th>Germany</th>
<th>United States</th>
<th>Japan</th>
<th>China</th>
</tr>
</thead>
<tbody>
<tr>
<td>1870</td>
<td>24.3</td>
<td>13.4</td>
<td>5.0</td>
<td>0.1</td>
<td>2.8</td>
</tr>
<tr>
<td>1913</td>
<td>18.5</td>
<td>18.0</td>
<td>9.0</td>
<td>0.8</td>
<td>2.0</td>
</tr>
<tr>
<td>1929</td>
<td>15.1</td>
<td>16.6</td>
<td>14.4</td>
<td>2.1</td>
<td>3.0</td>
</tr>
<tr>
<td>1950</td>
<td>10.2</td>
<td>3.9</td>
<td>16.2</td>
<td>1.3</td>
<td>0.9</td>
</tr>
<tr>
<td>1973</td>
<td>5.1</td>
<td>12.9</td>
<td>12.2</td>
<td>6.4</td>
<td>1.0</td>
</tr>
<tr>
<td>1990</td>
<td>5.3</td>
<td>12.0</td>
<td>11.3</td>
<td>8.2</td>
<td>1.8</td>
</tr>
<tr>
<td>2000</td>
<td>4.4</td>
<td>8.5</td>
<td>12.1</td>
<td>7.4</td>
<td>3.9</td>
</tr>
<tr>
<td>2012</td>
<td>2.6</td>
<td>7.7</td>
<td>8.4</td>
<td>4.4</td>
<td>11.2</td>
</tr>
<tr>
<td>2020 (projected)</td>
<td>1.9</td>
<td>5.3</td>
<td>8.8</td>
<td>3.9</td>
<td>12.1</td>
</tr>
<tr>
<td>2030 (projected)</td>
<td>1.4</td>
<td>3.6</td>
<td>7.3</td>
<td>3.2</td>
<td>15.0</td>
</tr>
</tbody>
</table>

Sources: Maddison 1995; UNCTAD various years; Subramanian 2011; and authors’ projections.

### Table 2.3 Exports and imports as percent of GDP in selected mega-traders (trade as percent of GDP)

<table>
<thead>
<tr>
<th>Item</th>
<th>Actual</th>
<th>Controlling for size</th>
<th>Controlling for size and income level</th>
<th>Controlling for size, income level, and oil-based economies</th>
</tr>
</thead>
<tbody>
<tr>
<td>United Kingdom 1870 (sample includes 26 countries)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exports</td>
<td>12.2</td>
<td>339.3***</td>
<td>84.0*</td>
<td>n.a</td>
</tr>
<tr>
<td>United States 1975 (sample includes 21 countries)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exports</td>
<td>8.5</td>
<td>–9.5</td>
<td>–37.0***</td>
<td>–36.3***</td>
</tr>
<tr>
<td>Imports</td>
<td>7.6</td>
<td>–30.5***</td>
<td>–37.7***</td>
<td>–37.5***</td>
</tr>
<tr>
<td>Total trade (exports and imports)</td>
<td>16.1</td>
<td>–20.9***</td>
<td>–35.5***</td>
<td>–35.1***</td>
</tr>
<tr>
<td>Japan 1990 (sample includes 131 countries)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exports</td>
<td>10.3</td>
<td>–33.8***</td>
<td>–56.8***</td>
<td>–55.6***</td>
</tr>
<tr>
<td>Imports</td>
<td>9.4</td>
<td>–44.3***</td>
<td>–49.4***</td>
<td>–51.4**</td>
</tr>
<tr>
<td>Total trade (exports and imports)</td>
<td>19.7</td>
<td>–40.4***</td>
<td>–52.9***</td>
<td>–53.7***</td>
</tr>
<tr>
<td>China 2008 (sample = 136 countries)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exports</td>
<td>35</td>
<td>79.9***</td>
<td>68.6**</td>
<td>80.5***</td>
</tr>
<tr>
<td>Imports</td>
<td>27.3</td>
<td>45.7***</td>
<td>46.2***</td>
<td>38.0**</td>
</tr>
<tr>
<td>Total trade (exports and imports)</td>
<td>62.2</td>
<td>64.6***</td>
<td>60.8***</td>
<td>62.0**</td>
</tr>
</tbody>
</table>

Note: All coefficients were obtained by running a regression of exports, imports and trade on variables indicated in column heads, plus a dummy for the country in question. The level of over-/under-trading is exp(dummy coefficient) – 1. A negative value denotes under-trading. * = significant at the 10 percent level, ** = significant at the 5 percent level, *** = significant at the 1 percent level.

Sources: Maddison for United Kingdom; IMF various years and Penn World Table 7.1 for all other countries.
Figure 2.6 Number of new signed preferential trade agreements, 1958–2012

Note: The year of the count is the year the World Trade Organization (WTO) was notified of the agreement. To simplify the classification of agreements, all agreements that are both economic integration agreements and customs unions or partial scope agreements are included in the “economic integration agreement” category.

Source: WTO 2011.
Table 2.4  Number and type of preferential trade agreements

<table>
<thead>
<tr>
<th>Type of agreement</th>
<th>Pre-WTO</th>
<th>1995–2000</th>
<th>Post-2000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WTO+ issues</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Customs</td>
<td>13</td>
<td>11</td>
<td>56</td>
</tr>
<tr>
<td>Antidumping</td>
<td>12</td>
<td>8</td>
<td>53</td>
</tr>
<tr>
<td>Countervailing measures</td>
<td>4</td>
<td>5</td>
<td>52</td>
</tr>
<tr>
<td>Export taxes</td>
<td>8</td>
<td>8</td>
<td>41</td>
</tr>
<tr>
<td>State aid</td>
<td>10</td>
<td>9</td>
<td>34</td>
</tr>
<tr>
<td>Trade-related intellectual property rights</td>
<td>6</td>
<td>4</td>
<td>41</td>
</tr>
<tr>
<td>Services</td>
<td>7</td>
<td>2</td>
<td>39</td>
</tr>
<tr>
<td>State trading enterprises</td>
<td>5</td>
<td>3</td>
<td>35</td>
</tr>
<tr>
<td>Technical barriers to trade</td>
<td>2</td>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td>Sanitary and phytosanitary standards</td>
<td>2</td>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>Public procurement</td>
<td>5</td>
<td>0</td>
<td>32</td>
</tr>
<tr>
<td>Trade-related investment measures</td>
<td>6</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td><strong>WTOX issues</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Competition policy</td>
<td>11</td>
<td>9</td>
<td>19</td>
</tr>
<tr>
<td>Movement of capital</td>
<td>6</td>
<td>5</td>
<td>38</td>
</tr>
<tr>
<td>Intellectual property rights</td>
<td>5</td>
<td>2</td>
<td>39</td>
</tr>
<tr>
<td>Investment</td>
<td>4</td>
<td>1</td>
<td>35</td>
</tr>
</tbody>
</table>

WTO = World Trade Organization

Note: WTO+ provisions concern commitments that already exist in WTO agreements but go beyond the WTO disciplines. WTOX provisions cover obligations that are outside the current WTO aegis.

Source: Baldwin 2011b.
Figure 2.7  Average most-favored nation (MFN) tariffs by income group, 1981–2009

Note: Spikes may reflect entry and exit of countries in the sample.

Figure 2.8  Index of services trade restrictiveness, by sector and region, 2008–10

The services trade restrictions index (STRI) at the regional level is calculated as a simple average of individual country's STRIs. The STRI in the cross-border air passenger transportation subsector comes from the QUASAR database of WTO (2007).

Regional abbreviations: HNO = High income non-OECD; SAR = South Asia; EAP = East Asia and Pacific; MENA = Middle East and North Africa; AFT = Sub-Saharan Africa; LAC = Latin America and Caribbean; ECA = Europe and Central Asia; OECD = High income OECD

Source: Borchert, Gootiz, and Mattoo 2012.
**Figure 2.9** Index of services trade restrictiveness and per capita GDP

services trade restrictiveness index (STRI)

Note: GDP per capita 2007, PPP (constant 2005 internat US dollars)

Source: Borchert, Gootiz, Mattoo 2012.

**Table 4.1** Policy responses to the challenges to globalization

<table>
<thead>
<tr>
<th>Level of response</th>
<th>Further liberalize</th>
<th>Maintain status quo</th>
<th>Retreat from globalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>National</td>
<td>In low-income countries, strengthen domestic supply capacity to exploit globalization.</td>
<td>Strengthen social insurance in high-income countries.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>In China and other middle-income countries, sustain growth to enable further liberalization.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>In high-income countries, revive growth and address “beleaguered middle class” and entitlements problems.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>International/collective</td>
<td>Prevent fragmentation and conflict.</td>
<td>Cooperate on taxation of mobile factors to sustain domestic safety net.</td>
<td>Create minimum safeguards to allow some trade protection?</td>
</tr>
<tr>
<td></td>
<td>Sustain multilateralism through a “China round.”</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Authors.
Figure 4.1  Tax rates on distributed corporate profits in selected OECD countries, 1981–2012

Note: Overall (corporate plus personal) tax rate on distributed profits are computed as effective statutory tax rates on distributions of domestic source income to a resident individual shareholder, taking account corporate income tax, personal income tax, and any type of integration or relief to reduce the effects of double taxation.

Source: OECD various years.
Figure 4.2  Share of capital in national income in selected OECD countries, 1980–2012

OECD = Organization for Economic Cooperation and Development.

Note: The share of capital is computed from the Annual Macroeconomic Database of the European Commission (AMECO) database using the adjusted wage share at current market prices.

Sources: Annual Macroeconomic Database of the European Commission (AMECO) various years.
Figure 4.3  Import shocks in the United States from Mexico, Japan, and China, 1962–2011

Note: Domestic absorption is GDP minus trade balance.

Sources: IMF, various years.
Table 4.2  Magnitude of import shocks to the United States from Japan, Mexico, and China

<table>
<thead>
<tr>
<th>Country</th>
<th>Period</th>
<th>Average</th>
<th>Change</th>
<th>Average</th>
<th>Change</th>
<th>China shock as multiple of earlier shocks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Without adjusting for per capita GDP</td>
</tr>
<tr>
<td>Japan</td>
<td>1970–90</td>
<td>373.6</td>
<td>355.2</td>
<td>6.79</td>
<td>6.15</td>
<td>1.2</td>
</tr>
<tr>
<td>Mexico</td>
<td>1980–2000</td>
<td>197.3</td>
<td>542.2</td>
<td>2.93</td>
<td>5.46</td>
<td>2.9</td>
</tr>
<tr>
<td>China</td>
<td>1990–2010</td>
<td>671.8</td>
<td>1258.6</td>
<td>8.49</td>
<td>14.92</td>
<td>n.a</td>
</tr>
</tbody>
</table>

n.a. = not applicable.

Note: Real imports are total nominal imports deflated by the unit price of imports. Import absorption is defined as (nominal) imports from each country divided by (nominal) domestic consumption (GDP less trade balance).

Sources: Authors, based on data from IMF various years; US Census Bureau various years; and Penn World Tables 7.1.
Figure 4.4  Relative income level of exporters to the European Union, Japan, and United States, 1980–2010

income level of partners as a share of own income level (percent)

Note: The measure represented here is the weighted average income level of exporters to the European Union, Japan, and the United States, excluding oil exporters (as defined by the International Monetary Fund) and small countries (countries with populations of less than 1 million). Income level is per capita GDP (purchasing power parity) using the rgdpch measure in the Penn World Tables. For example, if we call this index $R_{EU}$ for the European Union, it is computed as

$$R_{EU} = \frac{\sum_{i} (\frac{GDP_{EU}}{GDP_{i}}) \cdot (\frac{M_{i,EU}}{M_{EU}})}{\sum_{i} (\frac{GDP_{EU}}{GDP_{i}})}$$

where $M_{i,EU}$ is imports by the European Union from $i$ and $M_{EU}$ is total imports by the European Union. $R_{JP}$ and $R_{US}$ are identically computed for Japan and the United States.

Sources: IMF various years; Penn World Tables 7.1.
Figure 4.5  Relative income level of exporters to the European Union, Japan, and the United States with fixed weights, 1980–2010

Income level of partners as a share of own income level (percent)

Note: See note to figure 4.4. GDP weights are fixed to their initial 1980 value. The fixed-weight index $R_{EU}^{FW}$ becomes:

$$R_{EU}^{FW} = \sum \left( \frac{GD_{P,EU}}{GD_{P,EU,1980}} \right) \left( \frac{N_{EU,EU}}{N_{EU,EU}} \right).$$

Sources: IMF various years; Penn World Tables 7.1.
Figure 5.1  Share of capital in national income in selected country groups, 1995–2008

share of national income (percent)

Note: For emerging market countries, “compensation of employees” in the National Account Statistics was divided by GDP to compute the labor share. When series change, the last available series is used and completed by interpolation with the others using growth rates. The weighted averages (weighted by GDP in current dollars and purchasing power parity) were computed for eight emerging market countries (Argentina, Brazil, China, India, Korea, Mexico, Russia, and South Africa).

Sources: Authors, based on data from UN various years, Annual Macroeconomic Database of the European Commission (AMECO), World Bank various years, and Penn World Tables 7.1.
APPENDIX A

MEASURING VALUE-ADDED TRADE

In national account systems, trade is measured in gross terms, meaning that an export from a country is counted at its full value, whether this value was produced in the country or imported in part from another country (via foreign purchases of intermediate goods). It is thus possible to count a good that crosses frontiers at multiple stage of its production several times.

This accounting lies in contrast to the way in which GDP is computed. It is measured in terms of value added: The value of the intermediates used in production is subtracted from the value of the final good.

This distinction between gross and value-added trade has assumed significance in the wake of the ongoing process of the slicing up of the value chain across national boundaries. Slicing up is not a new phenomenon, but its magnitude has accelerated sharply in recent years, increasing the importance of proper measurement. Recent attempts have been made to correct this discrepancy and measure a consistent index of value-added trade by linking trade data and input-output tables. The basic idea is to link sources and uses of goods and services to be able to trace to its origin the value added embodied in an exported good.

Johnson and Noguera (2012) use estimates for 42 countries since 1970; their table 7 is the source of the value-added trade data in figure 2.1 in this paper. They use the concept of value-added exports (VAX), the ratio of value added in the country and exported to total exports. Recent research has refined the understanding of value-added trade by distinguishing various stages of production and trade (Koopman, Wang, and Wei 2013). In this paper, for reasons of simplicity, we adopt the VAX approach. To calculate the value-added trade data presented in the tables and figures presented in this appendix, we used the publicly available World Input Output Tables (available at http://www.wiod.org/database/iot.htm). These data span 15 years (1995–2009) and include 40 countries (including 27 EU countries and large developed and emerging market economies as well as a “rest of the world” aggregate), which represent 85 to 90 percent of global GDP.

We faithfully follow the methodology described in Johnson and Noguera (2012) to obtain a measure of bilateral exports in terms of both domestic and foreign value added. These exports are combined in various ways to derive the numbers used in the appendix figures and tables.

The caveat is that without firm-level data, one has to make the assumption that the production function is homogeneous within a sector across exporting and nonexporting firms. This assumption is probably not accurate, as exporters generally differ in size, productivity, and technology. The value-added trade data presented, although improvements over gross trade data, should be seen as first and necessarily imperfect approximations to the “real” value-added data.

The World Input-Output Table data are based on preliminary estimates that have since been revised, creating creates discrepancies for China in 2008 and 2009 and for India to a lesser extent. Data for the two years should therefore be used with caution (the problems with these years led us to use 2007 as the last year in the appendix tables).
### Table A.1  Mega-exporters, based on value-added trade (percent)

<table>
<thead>
<tr>
<th>Country/measure</th>
<th>Trade as a share of GDP</th>
<th>Overtrading, controlling for size</th>
<th>Overtrading, controlling for size and per capita GDP</th>
<th>Overtrading, controlling for size and per capita GDP and oil/small country dummies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japan (1995) (36 countries in sample)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gross exports</td>
<td>9.2</td>
<td>−55.8***</td>
<td>−53.8***</td>
<td>−52.8***</td>
</tr>
<tr>
<td>Value-add exports</td>
<td>8.5</td>
<td>−46.9***</td>
<td>−45.5***</td>
<td>−44.3**</td>
</tr>
<tr>
<td>China (2007) (35 countries in sample)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gross exports</td>
<td>37.9</td>
<td>160.7***</td>
<td>155.0***</td>
<td>161.0***</td>
</tr>
<tr>
<td>Value-add exports</td>
<td>27.7</td>
<td>103.2***</td>
<td>102.1***</td>
<td>109.6***</td>
</tr>
</tbody>
</table>

Note: See note to table 2.3; regressions are similar but use value-added exports and a restricted sample for comparability. * = significant at the 10 percent level, ** = significant at the 5 percent level, *** = significant at the 1 percent level.

Sources: Authors, based on data from World Input-Output Tables and Penn World Tables 7.1.

### Table A.2  China shock based on gross and value-added imports, 1995–2007

<table>
<thead>
<tr>
<th>Measure of imports</th>
<th>Real imports (dollars per working-age adults)</th>
<th>Import absorption (percent of domestic consumption)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Change</td>
</tr>
<tr>
<td>Gross</td>
<td>562.7</td>
<td>957.4</td>
</tr>
<tr>
<td>Value added</td>
<td>485.5</td>
<td>766.2</td>
</tr>
</tbody>
</table>

Note: See note to table 4.2.

Sources: Authors, based on data from World Input-Output Tables, Penn World Tables 7.1, and US Census Bureau.
Figure A.1  Gross and value-added imports by the United States as share of consumption, 1995–2009

share of domestic absorption (percent)

Note: See note to figure 4.3. Lines with (without) markers denote gross (value-added) imports.
Sources: Authors, based on data from World Input-Output Tables and IMF various years.
Figure A.2  Relative income level of exporters to the European Union, Japan, and the United States, based on value-added imports, 1995–2010

Note: See note to figure 4.4. Index here is similar but uses value-added import data.
Sources: Authors, based on World Input-Output Tables and Penn World Tables 7.1.

Figure A.3  Relative income level of exporters to the European Union, Japan, and the United States, based on value-added imports at fixed weights, 1980–2010

Note: See note to figure 4.5. Index here is similar but uses value-added import data.
Sources: Authors, based on World Input-Output Tables and Penn World Tables 7.1.
REFERENCES


The views expressed here are my own and are not necessarily those of Columbia University or the National Bureau of Economic Research. I thank participants in a seminar at the Federal Reserve Bank of San Francisco and the editors of the Journal of Economic Perspectives for their helpful comments. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2010 by Frederic S. Mishkin. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.
ABSTRACT

This paper examines what transformed a significant, but relatively mild, financial disruption into a full-fledged financial crisis. It discusses why, although the Lehman Brothers bankruptcy was a key trigger for the global financial crisis, three other events were at least as important: the AIG collapse on September 16, 2008; the run on the Reserve Primary Fund on the same day; and the struggle to get the Troubled Asset Relief Plan (TARP) plan approved by Congress over the following couple of weeks. The paper then looks at the policy responses to the financial crisis to evaluate whether they helped avoid a worldwide depression. The paper ends by discussing the policy challenges raised in the aftermath of the crisis.
The financial crisis of 2007 to 2009 can be divided into two distinct phases. The first, more limited, phase from August of 2007 to August of 2008 stemmed from losses in one, relatively small segment of the U.S. financial system—namely, subprime residential mortgages. Despite this disruption to financial markets, real GDP in the United States continued to rise into the second quarter of 2008, and forecasters were predicting only a mild recession. For example, the Congressional Budget Office (2008) released one of its periodic “The Budget and Economic Outlook: An Update” reports on September 8, 2008. It wrote: “According to CBO’s updated forecast for the rest of 2008 and for 2009, the economy is about halfway through an extended period of very slow growth. … Whether or not that period of slow growth will ultimately be designated a recession is still uncertain. However, the increase in the unemployment rate and the pace of economic growth are similar to conditions during previous mild recessions.” In keeping with that view, CBO projected that unemployment would rise modestly from 5.4 percent in 2008 to 6.2 percent in 2009, and that fourth-quarter to fourth-quarter real GDP would grow only 0.9 percent in 2008, but would rebound modestly to 1.8 percent growth in 2009. In summer of 2008, when I was serving on the Federal Reserve Board of Governors, there was even talk that the Fed might need to raise interest rates to keep inflation under control.

In mid-September 2008, however, the financial crisis entered a far more virulent phase. In rapid succession, the investment bank Lehman Brothers entered bankruptcy on September 15, 2008, the insurance firm AIG collapsed on September 16, 2008; there was a run on the Reserve Primary Fund money market fund on the same day; and the highly publicized struggle to pass the Troubled Asset Relief Program (TARP) began.

What caused the transformation from what appeared in mid-2008 to be a significant but fairly mild financial disruption into a full-fledged global financial crisis? Did the government responses to the global financial crisis help avoid a worldwide depression? What challenges do these government interventions raise for the world financial system and the economy going forward? Let’s start with a brief step back to the first phase of the global financial crisis.
THE FIRST PHASE: THE SUBPRIME MORTGAGE CRISIS

The first disruption of credit markets in the recent financial crisis is often dated to August 7, 2007, when the French bank BNP Paribas suspended redemption of shares held in some of its money market funds. A boom in U.S. housing prices had peaked around 2005. As housing prices started to decline, mortgage-backed financial securities—in many cases, securities based on subprime residential mortgages but then divided into more senior claims that were supposedly safe and junior claims that were recognized to be risky—began to experience huge losses. By early 2008, losses on these securities were estimated to be on the order of $500 billion dollars (for example, Greenlaw, Hatzius, Kashyap, and Shin (2008).

What developed in late 2007 and into 2008 was a series of runs on financial institutions, but instead of the classic bank run, it was, as described by Gorton and Metrick (2009), a run on the shadow banking system. A bank has deposits that are short-term liabilities and assets that are long-term loans. Thus, in a classic bank run, when bank depositors run to withdraw deposits, the bank cannot readily convert its long-term assets into cash. In the shadow banking system, institutions has short-term liabilities in the form of short-term borrowing, like repurchase agreements (or repos), which use longer-term assets like mortgage-backed securities as collateral. A key element of this borrowing is the use of a “haircut,” that is, a requirement that borrowers post collateral that is valued at more than the loan. For example, if a borrower took out a $100 million loan in a repo agreement, it might have to post $105 million of mortgage-backed securities as collateral, and the haircut would then be 5 percent. As the value of mortgage-backed securities fell and uncertainty about their future value increased, haircuts to levels as high as 50 percent. The result was that the same amount of collateral would now support less borrowing, leading to deleveraging in which financial institutions had to sell off assets. The resulting “fire sale” dynamic (discussed by Shleifer and Vishny in this issue) led to an adverse feedback loop in which the decline in asset values lowered the collateral’s value while further raising uncertainty, causing haircuts to rise further, which forced financial institutions to deleverage and sell more assets, and so on.

One signal of the resulting credit market disruptions appears in the interest rate spreads between safe and risky financial instruments. For example, the “TED spread” is the spread
between the interest rate on interbank lending (as measured by the LIBOR interest rate on three-month eurodollar deposits) and the interest rate on three-month U.S. Treasury bills. The TED spread provides an assessment of counterparty risk from one bank lending to another, reflecting both liquidity and credit risk concerns. Figure 1 shows how the TED spread rocketed up from an average of around 40 basis points (0.40 percentage points) before August 7, 2007, to 240 basis points by August 20, 2007, before abating somewhat.

The collapse of Bear Stearns in March 2008 was the most visible of these runs on the shadow financing system. Short-term financing for Bear Stearns dried up. Its long-term assets could not quickly be turned into ready cash at a fair price, and without access to short-term funding, it could not continue. The Federal Reserve brokered a deal for J.P. Morgan/Chase to purchase Bear, which was not unprecedented, but as part of the deal the Fed also took onto its books $30 billion of Bear Stearn’s’s toxic assets, which was unprecedented. However, this deal and the opening of new Federal Reserve lending facilities to investment banks helped restore some calm to the market. The TED spread surged to over 200 basis points in March 2008, but then fell back below 100 basis points.

By summer 2008, credit markets were clearly impaired and credit risk was rising, as can be seen by the rise in the spread between interest rates on Baa corporate bonds and Treasury bonds in Figure 1. However, the financial crisis looked like it could be contained. The Baa-Treasury spread had climbed to over 200 basis points, but these levels were similar to those that occurred in the aftermath of the mild recession in 2001. The TED spread, although elevated, was also below its peak values immediately after the revelations of problems at BNP Paribas and the Bear Stearns collapse. Many forecasters in the public and private hoped that the worst was over. After all, they reasoned that the subprime mortgage sector was only a small part of overall capital markets, and the losses in the related mortgage-backed securities, although substantial, seemed manageable. Indeed, the Congressional Budget Office (2008) was forecasting in early September 2008 that the Consumer Price Index would rise from 2.9 percent in 2007 to 4.7 percent in 2008. As discussed in Wessel (2009), there was talk in the Federal Reserve as to whether the easing phase of monetary policy might have to be reversed in order to contain inflation.
The story of this first phase of the 2007-2009 financial crisis has been discussed extensively in many places, including in symposia in the Winter 2009 and Winter 2010 issues of this journal. Here, the focus is on understanding what happened next.

THE SECOND PHASE: GLOBAL FINANCIAL CRISIS

In the space of a few short weeks in the fall 2008, everything changed. On Monday, September 15, 2008, after suffering losses in the subprime market, Lehman Brothers, the fourth-largest investment bank by asset size with over $600 billion in assets and 25,000 employees, filed for bankruptcy—the largest bankruptcy filing in U.S. history. Conventional discussions of the evolution of the financial crisis often view the Lehman bankruptcy as the key event that morphed the subprime crisis into a virulent global financial crisis. Although the Lehman bankruptcy led a large increase in uncertainty and a wave of distressed selling of securities that caused a collapse in asset prices and a drying up of liquidity, I will argue that the collapse of Lehman was followed by three events that were at least as important in causing the subprime crisis to go global: the AIG collapse on September 16, 2008; the run on the Reserve Primary Fund on the same day; and the struggle to get the Troubled Asset Relief Plan (TARP) plan approved by Congress over the following couple of weeks.

In considering these events, it’s also important to remember that the financial system had been greatly weakened before September 2008 in ways that had not yet been fully recognized at that time. Just as a relatively small sound or vibration can trigger an avalanche, if the snow conditions have made the danger of such an avalanche high, it may be that with given the amount of systemic risk embedded in the financial system, some other stress or failure of a financial institution would also have revealed the fragility of the financial system—and then led to a chain reaction that could also have tipped the financial system over the cliff.
The Lehman Bankruptcy

Many commentators have argued that the Treasury and the Fed’s decision to allow Lehmann to go bankrupt was a colossal mistake that turned a mild financial disruption into a global financial crisis. With hindsight, it is hard to argue that allowing Lehman to go bankrupt was the right decision. But it’s useful to remember that at the time, there was a plausible case for letting Lehman go into bankruptcy.

First, in practical terms, the U.S. government or its regulatory authorities had no authority to put Lehman into a government conservatorship so it could keep functioning, as the Treasury was able to do with Fannie and Freddie Mac. Thus, the only possible solution was to broker a purchase of Lehman. Barclays was in discussions about buying Lehman, but British bank regulators were skeptical and the Fed refused to take more bad assets on to its balance sheet, as it had done with Bear Stearns. Barclays ended up buying parts of Lehman a week after it declared bankruptcy.

Second, the bailout of Bear Stearns had extended the government safety net outside the banking system to investment banks, and the U.S. Treasury and the Federal Reserve were concerned about increasing moral hazard incentives on the part of a wider set of financial institutions to take on excessive risk. Indeed, as we now know, Lehman was going to extraordinary efforts, including engaging in shady accounting practices, to hide its leverage, even after the financial crisis started in August 2007.\(^3\) Letting Lehman fail would serve as a warning to other financial firms that they needed to reign in their risk taking.

Third, it was an open secret in the financial markets and among government officials that if any of the major investment banks would run into trouble, Lehman would be at the top of the list. Lehman was among the most leveraged of the major investment banks; it was unwilling to raise capital; it had a poor reputation for risk management; and it had a high exposure to losses on subprime mortgages because it had large holdings of securities tied to valuations of these

\(^3\) As described in the Examiner’s Report for the United States Bankruptcy Court, Southern District of New York, Valukas (2009), Lehman Brothers used a repo transaction, referred to as Repo 105, to reduce net leverage by $50 billion when reporting earnings at Q1 2008 and Q2 2008. In this transaction, repos were treated as sales, rather than borrowings, thereby taking them off the books. In addition, Lehman did not report that only $2 billion of $40 billion of liquid assets were readily accessible.
mortgages on its books (McDonald 2009; Sorkin, 2009). Sorkin (2009) documents that immediately after the Bear Stearns bailout, the U.S. Treasury Secretary immediately turned his attention to Lehman because he thought it would be the next trouble spot. With Lehman’s vulnerability already well-known, it seemed that Lehman’s was a natural test case to provide an object lesson to market participants that they should take measures to protect themselves. Indeed, many of the derivative contracts with Lehman’s counterparties were unwound successfully after Lehman’s bankruptcy.

Finally, the financial system in mid-September 2008 was far more vulnerable than almost all policymakers and market participants realized at that time. There is a distinct possibility that the financial system would have imploded even if Lehman had been bailed out.

*The AIG Collapse*

The Financial Products Unit of American International Group (AIG) had written over $400 billion dollars of insurance contracts called credit default swaps, which had to make payments when subprime mortgage securities suffered losses. With the Lehman Brothers collapse, it seemed more likely that the AIG might have to make enormous payments under these contracts, so short-term funding to AIG dried up. On September 16, 2008, the Federal Reserve stepped in with an $85 billion loan to keep AIG afloat (with total loans from the Fed and the U.S. government eventually rising to over $170 billion).

The enormous risk taking at AIG and its potential to blow up the financial system had been largely unrecognized by government officials, regulators, and markets. Once Bear Stearns had to be bailed out, it became apparent that a wider group of financial institutions could pose major systemic risks to the financial system. But in discussions at that time among regulators and academics about the need to regulate a wider group of financial institutions (in which I participated), AIG was *not* mentioned in the category of firms that would require special supervisory attention. This, along with Chairman Bernanke’s later statement in Congressional testimony about how angry he was that AIG took on such risk, describing AIG as effectively running a huge hedge fund inside an insurance company (Torres and Son, 2009), indicates how much the AIG blow up was a surprise.
Reserve Primary Fund

The same day of the AIG collapse—September 16, 2008—also saw a run on the Reserve Primary Fund, a large money mutual market fund run by Bruce Bent, one of the originators of money market mutual funds in 1970. Before the crisis, Bent had publicly criticized the industry for taking on too much risk in its asset holdings. He stated in a letter to the Securities and Exchange Commission in September 2007 (Bent, 2007): “When I first created the money market fund in 1970, it was designed with the tenets of safety and liquidity.” He added that these principles had “fallen by the wayside as portfolio managers chased the highest yield and compromised the integrity of the money fund.” Alas, Bent did not follow his own advice, and the Reserve Primary Fund held $785 million of Lehman paper. With the Lehman bankruptcy, the fund could no longer afford to redeem its shares at the par value of $1—a situation known as “breaking the buck”—and shareholders pulled out their money, with the fund losing 90 percent of its assets. A run on money market funds followed, with assets in institutional money market mutual funds falling from $1.36 trillion to $0.97 trillion from September to October 2008. In turn, this run put pressure on the banks, since a significant amount of bank funding was coming from bank commercial paper and certificates of deposits held by money market mutual funds.

TARP

In the wake of these events, U.S. Treasury Secretary, Hank Paulson, then proposed on September 19, 2008, the Troubled Asset Relief Program (TARP) in an infamous three-page document. In its original form, it would have given the U.S. Treasury the authorization, with no accountability to the Congress, to spend $700 billion purchasing subprime mortgage assets from troubled financial institutions, but which subsequently was used to inject capital into banking institutions. It soon became clear that Congress would vote down the original bill, which it did on September 29. Eventually the bill was finally passed on October 3, but passage required numerous “Christmas-tree” provisions such as a tax break for makers of toy wooden arrows.
The Broader Context

If the Federal Reserve had cut a deal with Barclays to rescue Lehman before bankruptcy, would the crisis have been defused? The underlying stresses in the financial system were all too real. A counterfactual history would have to take into account that a weakened Lehman, purchased before bankruptcy, might have later brought down Barclays. Rescuing Lehman would have increased moral hazard among other financial institutions, perhaps setting up a larger crash later. The costs of the AIG credit default swaps were eventually going to come due, quite possibly unexpectedly. Runs on various shadow banking institutions, like the run on Reserve Primary Fund and then on money market funds in general, were becoming more common. Here, rather than try to lay out a persuasive counterfactual history, I will emphasize two major changes that occurred by late September 2008.

First, even though markets had been digesting bad news about mortgage-backed securities since mid-2007, the events of September 2008 showed that risk taking was far more extensive than markets had realized and the fragility of the financial system was far greater than most market participants could have imagined. The AIG blow up and the run on the Reserve Primary Fund revealed that the financial system was engaged in what could be described as one huge “carry trade”. Technically carry trades are ones in which a trader borrows at a low interest rate to fund the purchase of assets that yield a high interest rate. Carry trades generate immediate profits, but may be very risky because the higher interest rate on the purchased assets may just reflect greater tail risk for that asset. AIG’s issuing of credit default swaps is a classic example of a type of carry trade, because the firm was earning large profits on the premiums paid on these contracts until the tail risk became a realization. In a prescient and now-famous paper, Rajan (2005) warned that this carry-trade problem was a danger to the financial system because incentives in compensation schemes for financial firms were leading to financial market participants engaging in financial transactions that produced immediate income, but exposed the financial system to massive risks.

Second, although markets had been watching government agencies scramble to deal with the financial crisis since late 2007, the events of September 2008 raised serious doubts that the U.S. government had the capability to manage the crisis. After all, the Fed and the U.S. Treasury proved unable to craft a solution so that Lehman would not fail. The AIG bailout was huge and
unexpected. TARP was originally proposed as a flimsy, three-page proposal, which raised concerns that the Treasury was unprepared, and the initial TARP proposal failed on a bipartisan vote. Even though the TARP legislation was eventually passed, the reputational damage was done.

After September 2008, the pattern of runs on the shadow banking system intensified and worsened. Banks began to horde cash and were unwilling to lend to each other, despite huge injections of liquidity into the financial system by the European Central Bank, the Bank of England and the Federal Reserve, The subprime crisis had become a full-fledged, global financial crisis.

The patterns of credit spreads tell the story. As shown in Figure 1, the TED spread rose from around 100 basis points during the week before the Lehman bankruptcy to over 300 basis points on September 17, the day after the liquidity squeeze on AIG and the Reserve Primary Fund materialized. The TED spread then dropped by 100 basis points, but as confidence in the ability and competence of the government to react quickly to contain the crisis weakened over the next couple of weeks, it climbed to over 450 basis points by October 10. The spread between interest rates on Baa corporate and Treasury bonds, shown in Figure 1, also rose by over 200 basis points and now rose well above levels that had been seen in 2001 during the prior recession period. The stock market crash also accelerated, with the week of October 6 showing the worst weekly decline in U.S. history.

Conditions in the financial markets continued to deteriorate. The public anger that resulted from the TARP “bailouts”— which involved injections of capital into financial institutions, with little restrictions on their use— became so intense that it became increasingly clear that the new Obama administration, taking office in January 2009, would not be able to get additional funds beyond those already allocated to TARP if needed. Figure 1 shows that although the TED spread fell from its peak in October 2008 with the help of government support to the financial sector, the spread between Baa and Treasury bonds continued to rise, peaking at over 500 basis points in December 2009. By the end of 2008, the stock market had fallen by over half from its peak in the fall of 2007.
The Links from Financial Crisis to Recession

Later data showed that the U.S. economy had turned down in the third quarter of 2008, falling at a -1.3 percent annual rate, but the recession that started in December 2007 became the worst economic contraction in the United States since World War II. Real U.S. GDP contracted sharply in the fourth quarter of 2008 and the first quarter of 2009, declining at annual rates of -5.4 and -6.4 percent, respectively. The unemployment rate skyrocketed, exceeding 10 percent by October 2009. A worldwide recession ensued as well. World economic growth fell at an annual rate of -6.4 percent in the fourth quarter of 2008 and -7.3 percent in the first quarter of 2009. A more extensive description of how financial crises lead to sharp economic downturns can be found in Mishkin (2011), but the basic story has three interrelated parts.

First, a financial crisis widens credit spreads, like the difference between interest rates on Baa corporate and Treasury bonds shown earlier in Figure 1. The result is that conventional monetary policy is defanged: even if interest rates on Treasury bonds fall because of a weakening economy and easing of monetary policy, the interest rates relevant to household and business purchase decisions go up, causing a drop in aggregate demand. Panel (a) of Figure 2 shows that Baa corporate bond rates barely budged at the beginning of the financial crisis in 2007 or during the Bear Stearns episode in March 2008, but climbed substantially in September 2008.

Second, the decline in asset prices during a financial crisis causes a decline in the value of collateral, which makes it harder for nonfinancial firms to borrow. In addition, the deterioration of balance sheets at financial firms, which have the expertise to mitigate adverse selection and moral hazard problems, causes their lending to fall, a process which is described by the term “deleveraging”, which causes spending to decline. Panel (b) of Figure 2 shows how total bank lending continued to rise early in the financial crisis in 2007, and even remained stable through March 2008 and the Bear Stearns rescue. Right after September 2008, bank lending rises largely because lenders were drawing heavily on already-established lines of credit, but by mid-2009 bank lending is on a downward trend. Of course, this decline should not only be attributed to the decline in the supply of loans, but also to the decline in the demand for loans as a result of weakening economic conditions.
Third, the general rise in uncertainty that occurs during a financial crisis also leads to an increase in asymmetric information, further hindering the ability of financial markets to allocate funds to households and businesses with productive investment opportunities. Panel (c) of Figure 2 shows how the market for asset-backed commercial paper, which had seemed to be recovering in mid-2008, which dwindled from daily average issuance of $64 billion at the beginning of September 2008 to $16.6 billion by the end of 2009.

**Policy Responses To The Financial Crisis**

The most powerful U.S. policy responses to the financial crisis came through policies that applied to the financial and banking system: conventional and unconventional monetary policies, bank “stress tests,” and bailouts of some banks and financial institutions. Many of these policies were implemented by the Federal Reserve, but others involved cooperation with fiscal authorities.

**Nonconventional Monetary Policy**

In 2002, when Ben Bernanke was member of the Board of Governors of the Federal Reserve, gave a speech (Bernanke, 2002) on the occasion of Milton Friedman’s 90th birthday and concluded by saying: “Regarding the Great Depression. You’re [referring to Milton Friedman and Anna Schwarz] right, we did it. We’re very sorry. But thanks to you, we won’t do it again.” He clearly meant it. The Federal Reserve’s modus operandi during the financial crisis can be characterized by saying that the Fed was engaged in massive experimentation in an unprecedented situation: that is, it was employing a large number of measures to contain the crisis, not knowing exactly which ones would work.

To be sure, the Fed started off using conventional monetary policy—that is, targeting a lower federal funds interest rate. Starting in the September 2007 meeting, the Federal Reserve lowered its federal funds rate target by ½ percentage point from 5.25% to 4.75%, and subsequently pushed the rate steadily downward. By April 2008, the rate was down to 2 percent, and by December 2008, the target range for the federal funds rate was 0 to 0.25%. Even before
the zero-bound for interest rate was reached, the Fed had turned to nonconventional monetary policy measures. Two nonconventional policy measures, liquidity provision asset purchases, result in an expansion of the central bank balance sheet and are therefore usually described under the heading “quantitative easing.” One other nonconventional measure is management of expectations. I will say a few words about each.

The first nonconventional form of monetary policy, *liquidity provision*, involves expanding Fed lending to both banks and other financial institutions. Liquidity provision is directed at maintaining the smooth functioning of financial markets, but it does affect household and business spending.

The traditional method for the Fed to provide liquidity has been through loans made at the discount rate—the interest rate on loans it makes to banks. In mid-August 2007, the Fed lowered the discount rate to 50 basis points (0.5 percentage points) above the federal funds rate target from the normal 100 basis points. It then lowered it further in March 2008 to only 25 basis points above the federal funds rate target. In addition, the Fed expanded the types of securities that would be eligible to be used as collateral. But discount lending has two problems: 1) it’s typically viewed as a bad signal for banks to borrow through the discount mechanism, because it suggests they had nowhere else to turn; and 2) discount lending has traditionally only gone to banks, not to other financial institutions.

To solve the problem of negative signals, the Fed set up a temporary Term Auction Facility (TAF) which enabled banks to borrow anonymously at a rate determined through a competitive auction. The TAF auctions started at amounts of $20 billion, but as the crisis worsened, the total loans outstanding though this mechanism rose to exceed $400 billion. (The European Central Bank conducted similar operations, with one auction in June 2008 leading to lending of over 400 billion euros.)

The Fed also invented new lending programs to broaden its provision of liquidity to beyond banking institutions. These included lending to investment banks, and lending to promote purchases of commercial paper, mortgage backed-securities and other asset-backed
securities. In addition, the Fed engaged in lending to prop up Bear Stearns, AIG and to Fannie Mae and Freddie Mac. The enlargement of the Fed’s lending programs during the 2007-2009 period was remarkable, expanding the Fed’s balance sheet by over $1 trillion by the end of 2008, with the balance-sheet expansion continuing into 2009. The number of new programs over the course of the crisis spawned a whole new set of acronyms: TSLF, PDCF, AMLF, MMIFF, CPFF and TALF.

Yet another method to increase liquidity was through swap lines with foreign central banks. These foreign central banks also engaged in massive amounts of liquidity provision, but could create liquidity only in their own domestic currency, while many of their financial institutions required dollar funding to conduct their operations. The Federal Reserve provided foreign central banks with U.S. dollar deposits in exchange for deposits in their home currency, often in essentially unlimited amounts.

Overall, the available research suggests that liquidity provision did help stabilize financial markets during this crisis. For example, McAndrews, Sarkar and Wang (2008) find that announcements about the Term Auction Facility (TAF) did significantly lower credit spreads. Wu (2008), Christensen, Lopez and Rudebusch (2009) and Sarkar and Shrader (2010) also conclude that the TAF and other credit facilities helped lower interest rates. Baba and Packer (2009) and McAndrews (2009), Goldberg, Kennedy and Miu (2010) find that the U.S. dollar swap facilities helped improve the performance of the dollar swap markets. Using a similar event-study methodology, Ait-Sahalia et al. (2010) find that liquidity provision in not only the United States, but also in the United Kingdom and Japan, did help lower interbank risk premiums.

---

6 The lending to J.P. Morgan to prop up Bear Stearns was in effect a purchase of assets. In order for the Federal Reserve to abide by its legal authority, it could not purchase private assets outright. Instead, it made a nonrecourse loan: that is, the Fed had no recourse to require J.P. Morgan to pay back the loan, but instead would take ownership of the collateral, the $30 billion of toxic assets. Hence the Fed would bear any losses or gains on these assets so in economic terms it had purchased these assets.

1 For an alternative view, Taylor and Williams (2009) find no evidence that actual lending from the Term Auction Facility (TAF) helped to ease credit markets.
The second category of nonconventional approaches to monetary policy, asset purchases, is based on the belief that a direct purchase can stimulate spending by raising prices on particular classes of bonds, thereby lowering the interest rates that households and businesses have to pay. This policy began with the purchase of $300 billion of long-term Treasury bonds, which started in March 2009 and ended in October 2009. Empirical evidence on a previous Fed attempt to lower long-term interest rates relative to short-term rates in the 1960s—which was dubbed “Operation Twist” because its intent was to “twist” and flatten the yield curve—deemed that it didn’t work (Modigliani and Sutch, 1967). However, as Solow and Tobin (1987) pointed out, Federal Reserve purchases of long-term Treasury bonds were small and ended up being offset by issuance of long-term bonds.

The more important asset purchase program, announced in November 2008 and terminated in March 2010, was the outright purchase of $1.25 trillion of mortgage-backed securities. The Fed purchased mortgage-backed securities in an attempt to lower residential mortgage rates, thereby stimulating the demand for housing, which would not only stabilize housing prices and the financial markets, but would also stimulate spending on residential construction. Research on the impact of the Fed’s large-scale asset purchases during the global financial crisis by Gagnon, Raskin, Remache and Sack (2010), finds that these programs lowered long-term bond rates relative to short rates on the order of 50 basis points, and lowered interest rates on mortgage-backed securities even further by improving liquidity in this market, thereby having a substantial impact on residential mortgage rates.

Both liquidity provision and asset purchased fit under the general heading of quantitative easing—that is, ways in which the Federal Reserve greatly expanded the monetary base along with its balance sheet. There has been some question as to whether this expansion of the monetary base, by itself, could stimulate the economy. It’s not clear why this would work: as Curdia and Woodford (2010) argue, why should an expansion of the monetary base lead to higher aggregate demand when it was unable to further lower interest rates or stimulate bank lending? In addition, evidence from the Japanese episode does not provide much support that a pure expansion of a central bank’s balance sheet is particularly effective in stimulating aggregate demand (Kuttner, 2004). Bernanke (2009) has also expressed his skepticism that quantitative easing, by itself would be effective. He indicated that the expansion of the balance sheet should
instead be viewed as a result of what he referred to as *credit easing*, that is, an attempt to lower spreads between different asset classes through asset purchases and liquidity provision.

A final nonconventional approach to monetary policy is *management of expectations*. In March 2009, the Fed Open Market Committee added to its policy statement that it would maintain “exceptionally low” interest rates “for an extended period. There is theoretical support for the proposition that a commitment to keep short-term interest rates low for a substantial period of time helps lower long-term interest rates and also raises inflation expectations, thereby reducing the real interest rate (Eggertsson and Woodford, 2003, 2004; Woodford, 2003). But at least so far, no empirical evidence is available for how effective management of expectations was during this episode.

The Bank Stress Tests

Financial markets began to recover in the first half of 2009. The provision of huge amounts of liquidity appeared to do the trick in the interbank lending market, with the TED spread falling from its peak of over 400 basis points in October to below 100 basis points in January 2009. This spread fell to below pre-crisis levels (less than 20 basis points) by May 2009. Credit spreads also began to fall with the Baa-Treasury spread declining from its peak in October, but at a slower pace than the improvement in the interbank market, as illustrated in Figure 1. By late 2009, however, credit spreads were returning to normal, reaching levels that were just a little above those before the crisis, and actually lower than the spreads that existed in 2002, shortly after the previous recession ended. The stock market also began to recover from its trough in March 2009, leading to a sustained bull market in which it rose over 50 percent over the next year.

A key element in the financial market recovery was the U.S. Treasury’s requirement, announced in February 2009, that the 19 largest banking institutions undergo the Supervisory Capital Assessment Program or SCAP, which are commonly referred to as *stress tests*. The stress tests were a supervisory assessment, led by the Federal Reserve in cooperation with the Office of the Comptroller of the Currency and the FDIC, of the balance sheet position of these banks to ensure that they had sufficient capital to withstand bad macroeconomic outcomes. The stress tests were designed as a forward-looking exercise to project possible erosion of bank
capital under two scenarios: the baseline consensus forecast by private sector economists as of February 2009, which entailed a continuing decline in economic activity, and a worse scenario of a much more severe recession (for details of the scenarios, see Board of Governors of the Federal Reserve System, 2009).

The Treasury announced the results in early May 2009 and they were well received by market participants, allowing these banks to raise substantial amounts of capital from private capital markets that were required by the stress tests. The stress tests were a key factor that helped increase the amount of information in the marketplace, thereby reducing asymmetric information and adverse selection and moral hazard problems. Hoshi and Kashyap (forthcoming) found that similar stress tests in Japan in 2003 were a key element of the recovery of the Japanese banking system after the “lost decade” from 1992 to 2002.

**Bailing Out Financial Institutions**

Some of the Fed’s liquidity provision was to bail out financial institution, as occurred with Bear Stearns, AIG and the government-sponsored enterprises Fannie Mae and Freddie Mac. In each of these cases, the Federal Reserve provided this liquidity in cooperation with the U.S. Treasury, which also made large loans.

Although the Troubled Asset Relief Plan (TARP) was initially intended to purchase subprime mortgage assets to help prop up financial institutions’ balance sheets, it soon became clear that agreeing on a prices for those assets was unworkable. The Treasury switched to using the TARP funds to inject capital into financial institutions, thereby shoring up their balance sheets more directly. In addition, on September 29, the U.S. Treasury announced a Temporary Guarantee Program for Money Market Funds, which insured that investors would receive at least the $1 par value per share. On October 14, 2008, the FDIC announced the Temporary Liquidity Guarantee Program (TLGP) that guaranteed newly-issued senior unsecured bank debt, such as federal funds (loans of deposits at the Federal Reserve) and commercial paper, as well as noninterest bearing transaction accounts. Its stated purpose was to “strengthen confidence and encourage liquidity in the banking system” (FDIC, 2008). Although these programs were initially intended to last less than a year, they have been extended several times.
The spreading bank failures in Europe in fall 2008 led to similar bailouts of financial institutions: for example, the U.K. Treasury set up a bailout plan that guaranteed 250 billion pounds of bank liabilities, added 100 billion pounds to a facility that swaps these assets for government bonds, and allowed the U.K. government to buy up to 50 billion pounds of equity stakes in British banks. Allessandri and Haldane (2009) discuss $10 trillion worth of these bailout packages across 20 countries, which includes both guaranteeing the debt of the banks and injecting capital into them. There was a high degree of international coordination in these policies.

Ait-Sahalia et al. (2010) find that comprehensive bailouts which helped recapitalize the financial sector did help lower interbank risk premiums, but bailouts of individual banks on an ad hoc basis were received poorly by the markets and led to a rise in interbank risk premiums. A plausible explanation is that when governments pursue ad hoc bailouts, it suggests to markets that the problem in the credit markets may be worse than they expected. In contrast, pursuing a comprehensive approach to recapitalize the financial system helps to restore confidence and to unfreeze the credit markets. Furthermore, they find that there were strong spillovers from actions taken in one country to others, suggesting the benefits of a coordinated policy response between countries to cope with a global financial crisis.

*Expansionary Fiscal Policy*

Fiscal stimulus to directly increase aggregate demand was another key piece of the government response to the global financial crisis, both in the United States and in many other countries. The incoming Obama administration pushed for the $787 billion fiscal stimulus package, the American Recovery and Reinvestment Act of 2009. The plan featured $288 billion of tax cuts and $499 billion in government spending increases. The evidence on the effect of the fiscal stimulus package is mixed, but two arguments suggest that it was far less important to addressing the financial crisis than were actions by central banks to provide liquidity and government recapitalization and guarantees of the financial system.

First, as a basic matter of timing, most of the additional government stimulus package did not come on line until late 2009 and into 2010. While one can construct a theoretical
argument that the expectation of the stimulus package helped to reassure financial markets, any
direct effect of the stimulus on the financial crisis through the early months of 2009 was
necessarily quite limited.

Second, there has is a very active debate about how much a fiscal stimulus will affect
output. For example, Hall (2009) summarizes the theoretical and empirical evidence on fiscal
stimulus as yielding an output multiplier between 0.7 and 1. Analysis of the stimulus using
dynamic stochastic general equilibrium models in Cogan et al. (2009) and Uhlig (2010) find that
multipliers in the 0.6 range, while econometric evidence employed by Barro and Redlick (2009)
finds an output multiplier of around 0.7. On the other hand, economists in the Obama
Administration and the Congressional Budget Office, using Keynesian, large scale econometric
models, estimated that the output multiplier for fiscal stimulus were well above one. Moreover,
a number of models point out that when the interest rate falls to the zero lower bound after a
large negative aggregate demand shock, the output multiplier from a fiscal stimulus may be
much higher because expansionary fiscal policy raises inflation expectations, thereby lowering
real interest rates: for example, Eggertsson (2009), Woodford (2010) and Christiano,
Eichenbaum and Rebelo (2009) find such a result in calibrated New Keynesian models. The
controversy over discretionary fiscal stimulus is explored by Auerbach and Gale in this issue.

*Have Policies to Ameliorate the Financial Crisis Succeeded?*

The question of whether or in which ways the policies to defuse the financial crisis have
succeeded will be debated for years. One group of skeptics points out that households and firms
have seen tighter credit standards and a higher cost of credit during the recession, from which
they conclude that monetary policy has not been effective during the recent financial crisis (for
example, Krugman, 2008). Another view holds that many government actions were ineffective,
while others may have raise the perceived level of risk in financial markets (for example, see
Taylor, 2009).

My own view, as I have argued more extensively elsewhere (Mishkin, 2009), it that
conclusions about the effectiveness of policy should begin by considering the counterfactual—
that is, what would the likely course of events without the policy interventions. For example, if
the Federal Reserve had not lowered the federal funds rate by over 500 basis points starting in September 2007, its clear that, interest rates on default-free Treasury securities would have been higher, but I believe further that credit spreads would have widened by even more than they did during this crisis, because the weaker economy would have made conditions in financial markets even more stressed. The outcome would then surely have been that households and firms would have faced much higher interest rates, with the result that household and firm spending would have declined even more precipitously than we saw. The banking stress tests and systematic efforts to recapitalize the banking system also seem to have been useful. Some parts of the government intervention were less useful than others. But taken as a whole, I believe the government actions helped to prevent a far deeper recession and even possibly a depression.

AFTERMATH: CLEANING UP AFTER THE CRISIS

The global financial crisis of 2007-2009 appears to be waning. There are three key areas of government policies to clean up after the crisis in order to restore the world’s financial sector and the broader economy to health.

Shrinking Central Bank Balance Sheets

Actions by central banks to contain the global financial crisis resulted in huge expansions of their balance sheets. The expansion of balance sheets arising from liquidity provision is typically easy to reverse because most of the liquidity facilities have provided loans at interest rates that are higher than market rates during normal times. As financial markets return to normal, market participants are no longer willing to borrow at above-market ranks, this source of balance sheet expansion naturally reverses itself as the financial system recovers—which is exactly what has happened.

The asset market purchases of long-term mortgage-backed securities are not self-liquidating in this way. Over $1 trillion of the mortgage-backed securities have maturities of ten years or more. Thus, a strategy of just letting them run off will leave the Federal Reserve in this market for a long time, which raises several issues. First, by holding these securities the Federal
Reserve will be exposed to both credit and interest rate risk.\(^2\) Second, the presence of private securities on the Federal Reserve balance sheet means that the Fed has become directly involved in perhaps the most politicized financial market in the United States. The public and Congress may begin to hold the Fed accountable for what happens specifically to mortgage rates, rather than to interest rates in general. Politicians may tend to see the Fed as institutionally responsible for developments in the housing markets.

Can the Fed extricate itself from this situation by selling the mortgage-backed securities? The experience of the end of the purchase program for mortgage-backed securities at the end of March 2010 is encouraging. For some months before this date, the Fed had been in essence the sole buyer in this market. However, given that financial markets had stabilized and that the end of the purchase program was well publicized, the Fed’s exit from the market did not cause any disruption. The spreads of mortgage-backed securities over Treasury bills did not rise after April 1, 2010. This experience suggests that if the Fed announces a program of asset sales well in advance and financial markets are functioning normally, it should be able to liquidate its positions. Of course, if this turns out not to be the case, then the Fed could discontinue its sales and announce that its sales are contingent on the market continuing to function normally.

A final concern sometimes raised is that the expansion in the monetary base will necessarily be inflationary, but this is unlikely to be the case in the current environment. The reason is that banks are perfectly happy to hold huge amounts of excess reserves—thus essentially neutralizing the effect this money would have on demand or the price level—as long as they are paid interest on the reserves, as is now the case. However, purchase of long-term government bonds has raised concerns that the Fed is willing to accommodate profligate fiscal policy by monetizing government debt, and this does have the potential to unanchor inflation expectations, which could have inflationary consequences in the future.

---

\(^2\) Because the balance sheet of the Federal Reserve is in effect part of the overall government balance sheet, it is not clear why potential losses on the Fed’s balance sheet should matter. However, such losses would likely result in severe criticisms of the Federal Reserve and so weaken its independence.
Too-Big-To-Fail

The global financial crisis has encouraged efforts to revamp financial regulation. French et. al. (2010) offer discussions of financial regulation in the future, as do some of the other papers in this symposium. Here I will just focus on one issue, the too-big-to-fail problem.

Too-big-to-fail is a misnomer. A financial firm can be systemically important—that is, its failure can threaten the health of the financial system—either because it is so large or its activities are so interconnected with the rest of the financial system. A more accurate term would be too-interconnected-to-fail or too-systemically-important-to-fail. The failure of the hedge fund Long Term Capital Management in 1998 is a classic example of a firm that was not enormous in size, but was systemically important. In any case, the difficulty is that when creditors know that a firm falls into this category, they can expect government provide some assistance if the firm gets into trouble, which means that they have less incentive to monitor the firm and pull out their money if it is taking on too much risk. Of course, this makes excessive risk-taking more likely, and raises the cost to taxpayers of the eventual government bailout.

Too-big-to-fail is now a larger problem than before, in part because banks have merged in a way that creates even larger banking institutions, and because with the Fed bailout of Bear Stearns in March 2008, and then the financial assistance to AIG by the Fed and the U.S. Treasury in September of 2008, it has become clear that a much wider range of financial firms are likely to be considered to be too-big-to-fail in the future. Indeed, the most prominent case of a firm that was not bailed out—Lehman Brothers in September 2008—was followed by such a severe crisis, that it is unlikely that governments would let this happen again. In the wake of the Lehman failure, governments throughout the world bailed out or guaranteed all their major financial institutions.

One way to address the too-big-to-fail problem is to limit the size of financial institutions, which might involve either the breakup of large financial institutions and/or limits on what activities banking institutions can engage. However, arbitrary limits on their size or activities might well decrease the efficiency or raise other risks in the financial system. An alternative view is to subject systemically important institutions to greater regulatory oversight, say by a systemic regulator (as discussed in Mishkin, 2010a; French et. al., 2010), or by imposing larger capital requirements for systemically important financial firms.
The Dodd-Frank financial reform bill passed in summer 2010 gives the federal government one more tool for dealing with systemically important financial companies. Before Dodd-Frank, the U.S. government only could take over individual banking institutions, but not financial holding companies that own banks and other financial institutions. (In other words, it could take over Citibank, but not Citigroup or a free-standing investment bank like Lehman Brothers.) It used to be that the government had only two alternatives with such firms: send them into bankruptcy or bail them out. Now, the federal government has “resolution authority” over such firms, which means that they can treat them as they would an insolvent bank. Critics have expressed concerns that this federal resolution authority will further entrench too-big-to-fail and so make the moral hazard problem worse (for example, Wallison, 2010). As with all regulatory authority, the devil will be in the details. But the new resolution authority is likely to help limit moral hazard because it gives the government a big stick to force systemically important financial institutions to desist from risk taking or to raise more capital—or else to face a government takeover that imposes costs on managers and shareholders.

Retrenching Fiscal Policy

The combination of massive bailouts, fiscal stimulus packages, and the sharp economic contractions that reduced tax revenue have shifted the fiscal situation for many countries. As Reinhart and Rogoff (2009) point out, the aftermath of financial crises is almost always a large increase in government indebtedness and we have seen exactly this pattern in the aftermath of the current crisis. Budget deficits over 10 percent of GDP in advanced countries like the United States have been common in 2009 and 2010. This rise in government borrowing can even raise the risk of sovereign debt defaults, which can be a particular problem if sovereign debt is being held by many banks as a “safe” asset. This risk has become a serious concern in Europe after the Greek sovereign debt crisis.

As budget deficits surged after the crisis, the ratio of government debt to GDP is projected to jump to very high levels in many countries. In the next decade or so, getting fiscal houses in order will become one of the highest priorities for government policy throughout the world. In many countries, governments already faced a long-term problem of unsustainable
spending growth on health care and pensions; the current fiscal imbalances have brought those problems forward in time from the long-term into the middle-term, and in some countries into the short-term and the immediate future.

CONCLUSION

What started in 2007 as a crisis in one small part of the financial system led to a worldwide economic conflagration by late 2008 and early 2009. There are two key lessons from what has happened. First, the global financial system is far more interconnected than was previously recognized and excessive risk taking that threatened the collapse of the world financial system was far more pervasive than almost anyone realized. Understanding how systemic risk can arise and designing policies to rein in this risk taking are tasks of the highest priority. Second, extraordinary actions by central banks and governments have contained this global financial crisis, but successfully unwinding these policies will prove to be a highly challenging task.
REFERENCES


24


Eggertsson, Gauti B.. 2006. “Was the New Deal contractionary?” Federal Reserve Bank of New York Staff Report, No. 264

Egertsson, Gauti B. 2009. “What fiscal policy is effective at the zero interest rates.” Federal Reserve Bank of New York Staff Report, No. 402


Figures

Figure 1: Credit Spreads 2000-2009

Source: FRED, Federal Reserve Bank of St. Louis British Bankers’ Association

Note: TED spread is the difference between the 3-month Libor rate and the constant maturity 3-month Treasury bill rate. Baa spread is the difference between constant maturity Baa rate and the 10-year constant maturity Treasury bond rate.
Figure 2: Financial Variables, 2007-2009

a) Baa corporate bond rate
b) Bank Lending
(c) Asset-Backed Commercial Paper Issuance

*Source:* FRED database, Federal Reserve Bank of St. Louis, and Federal Reserve Board of Governors.

*Note:* The Baa corporate bond rate is the constant maturity interest rate, asset-back commercial paper issuance is the daily average of issuance of asset-backed commercial paper, and bank lending is total loans and leases of commercial banks.
What makes growth sustained?

Andrew Berg, Jonathan D. Ostry, Jeromin Zettelmeyer

International Monetary Fund, Washington, DC, USA
European Bank for Reconstruction and Development, London, UK

ARTICLE INFO

Article history:
Received 8 June 2010
Received in revised form 28 July 2011
Accepted 15 August 2011

JEL classifications:
F143
011
02

Keywords:
Growth
Accelerations
Structural breaks
Income inequality

ABSTRACT

We identify structural breaks in economic growth in 140 countries and use these to define “growth spells”: periods of high growth preceded by an upbreak and ending either with a downbreak or with the end of the sample. Growth spells tend to be shorter in African and Latin American countries than elsewhere. We find evidence that growth duration is positively related to: the degree of equality of the income distribution; democratic institutions; export orientation (with higher propensities to export manufactures, greater openness to FDI, and avoidance of exchange rate overvaluation favorable for duration); and macroeconomic stability.

© 2011 International Monetary Fund. Published by Elsevier B.V. All rights reserved.

1. Introduction

Perhaps the most important question confronting policymakers in low-income and emerging market countries is how to sustain economic growth. Until recently, however, the economics literature provided little guidance on this issue. To be sure, since the early 1990s, a body of work centered on cross-country growth regressions has aimed to explain differences in long-term growth between, say, the miracle episodes in Asia and the stagnation in sub-Saharan Africa and Latin America. However, this work ignored a fundamental property of growth: that growth looks like, perhaps so that other countries can take a different course for the worse, it would be useful to know what the path out of a downbreak is, and what is going on around the time of the transition, and during the growth episode, to uncover any useful patterns. Likewise, if a country has been growing well for a number of years, but suddenly changes course for the worse, it would be useful to know what the path out of growth looks like, perhaps so that other countries can take a different fork in the road. Papers that attempt to uncover the informational content of growth transitions—inspired by Pritchett (2000) and related work, such as Aguiar and Gopinath (2007), Ben-David and Papell (1998), and Easterly et al. (1993)—include: Hausmann et al. (2005), Hausmann et al. (2006), Jerzmanowski (2006), Jones and Olken (2008), Pattillo et al. (2005), Reddy and Minoui (2009) and Rodrik (1999). The results from this literature have been mixed. They partly confirm some of the elements that were thought to be important based on the cross-country approach, such as the importance of institutions. However, the papers also suggest that growth transitions remain largely a mystery, in that the “usual suspects” explain only a small perspective, of why some growth episodes tend to end more quickly and abruptly, or why some downturns may be relatively protracted.1

A more promising approach may involve exploiting the information in turning points in countries’ growth performance. If an economy has been falling off a cliff for a number of years and then turns itself around and starts climbing a mountain, it makes sense to ask what is going on around the time of the transition, and during the growth episode, to uncover any useful patterns. Likewise, if a country has been growing well for a number of years, but suddenly changes course for the worse, it would be useful to know what the path out of growth looks like, perhaps so that other countries can take a different fork in the road. Papers that attempt to uncover the informational content of growth transitions—inspired by Pritchett (2000) and related work, such as Aguiar and Gopinath (2007), Ben-David and Papell (1998), and Easterly et al. (1993)—include: Hausmann et al. (2005), Hausmann et al. (2006), Jerzmanowski (2006), Jones and Olken (2008), Pattillo et al. (2005), Reddy and Minoui (2009) and Rodrik (1999). The results from this literature have been mixed. They partly confirm some of the elements that were thought to be important based on the cross-country approach, such as the importance of institutions. However, the papers also suggest that growth transitions remain largely a mystery, in that the “usual suspects” explain only a small perspective, of why some growth episodes tend to end more quickly and abruptly, or why some downturns may be relatively protracted.1

1 Panel regressions shed some light on these issues, but may not capture turning points well and are misspecified if the growth dynamics are not captured in a stable linear relationship with a set of fundamentals.
fraction of what is going on during a transition. More specifically, currency depreciations and political regime changes seem to be correlated with growth accelerations (Hausmann et al., 2005), while collapses in investment play a role in downbreaks. Growth decelerations are found to be associated with macroeconomic instability, conflict, and export collapses (Hausmann et al., 2006). Hence, one tentative conclusion from this literature is that what matters for getting growth going may be different from what is important to keep it going.

The present paper contributes to this literature by focusing squarely on the second issue, namely the predictors of growth duration. Closing the per capita income gap with rich countries requires long periods of fast growth in the developing world. Surges in growth are in fact relatively common in the developing world, even in regions that have done very badly over the past few decades (e.g., sub-Saharan Africa). What really sets poor-performing regions apart is that their growth spells have tended to end relatively quickly (e.g., in comparison with East Asian or industrial countries). The question of how to forestall spells has tended to end relatively quickly (e.g., in comparison with East Asian or industrial countries). The question of how to forestall the end of growth spells is thus critical, especially for the large number of developing countries that has recently been enjoying strong growth.

We approach the topic somewhat differently than the literature that precedes us. Focusing on the before-and-after of a deceleration episode misses a great deal of potential information, because it does not tell us what a country (or its environment) was doing right prior to its deceleration. Studies that focus on deceleration events are not well-placed to draw lessons from the fact that some decelerations are preceded by much longer periods of high growth than others. We attempt to capture this information by moving the object of inquiry to duration per se: that is, by studying the determinants of the length of growth spells. We do so by applying duration analysis techniques that are common in medical or microeconomic applications (for example, studies that examine the length of unemployment spells). A further advantage is that the duration approach can easily take account of censored observations. It can exploit the information contained in growth spells that are still ongoing, inferring a protective relationship for a covariate that is higher in long continuing spells than in shorter spells that end in-sample.

The object of our analysis is the “growth spell:” the time period between a growth acceleration and a deceleration. To identify accelerations and decelerations, we combine both statistical structural break tests and economic criteria. Relying exclusively on ad hoc economic criteria may not be the best approach if year-to-year volatility in the underlying growth series differs substantially across countries, as is indeed the case; we would risk identifying too many spurious breaks, particularly in countries with higher growth volatility. But relying exclusively on statistically-determined structural breaks in growth may not be enough, because some statistically significant breaks in growth may be too small to be of much interest economically.

Having identified growth spells, we explore the potential correlates of spell duration by estimating a proportional hazard model with time-varying covariates. This model relates the probability that a growth spell will end to a variety of economic and political variables. In doing so, we distinguish between “initial conditions” in place at the time of an acceleration and changes that take place during a growth spell. The latter are particularly relevant for the question of what policies can extend the life of an ongoing growth spell.

Within the literature on growth transitions, our approach relates most closely to recent papers by Hausmann et al. (2006) and Jerzmanowski (2006). While Hausmann et al. (2006) also use duration analysis, they focus on the length of stagnations rather than that of growth spells. Another difference with respect to the present paper is that they identify stagnations using only an economic criterion, without reference to the question of whether these breaks are statistically significant. Jerzmanowski (2006) is easily the most ambitious paper in this literature, and perhaps the most faithful to the Pritchett (2000) idea that developing country growth can be classified into structurally different regimes. He estimates a Markov-switching model of growth with four such regimes: miracle growth; stable growth; stagnation; and crisis. These four regimes and the 16 transition probabilities among regimes are estimated simultaneously. However, the approach is so informationally demanding that it can examine only one potential determinant of transition probabilities at a time. In contrast, our paper identifies just two regimes using structural break analysis and then, in a second step, investigates the probability of a regime switch (from growth to stagnation) using duration analysis. In doing so, we study many potential factors influencing the probability that a growth spell might end.

Our main findings confirm some previous results in the literature: external shocks and macroeconomic volatility are negatively associated with the length of growth spells, while good political institutions predict longer growth spells. We also have some more surprising findings. Trade liberalization seems to not only be associated with getting growth going, as emphasized by previous authors, but also with sustaining it—particularly when combined with competitive exchange rates, current account surpluses, and an external capital structure weighted toward foreign domestic investment. Furthermore, we find that export composition matters. We find that the manufacturing share in exports, and more generally, export product sophistication tend to predict prolonged growth. Most strikingly, we find that the duration of growth spells is strongly related to income distribution: more equal societies tend to sustain growth longer. On the whole, these results share some of the flavor of recent work on the political economy of growth and development, as briefly discussed in Section 4 and in our Conclusion.

2 Structural breaks and “growth spells”

We apply a variant of a procedure proposed by Bai and Perron (1998, 2003) to test for multiple structural breaks in time series when both the total number and the location of breaks are unknown. Our approach differs from the Bai–Perron approach in that it uses sample-specific critical values that take into account heteroskedasticity and sample size as opposed to asymptotic critical values; and in that it extends Bai–Perron’s algorithm for sequential testing of structural breaks, as described below. Antoshin et al. (2008) describes these extensions in more detail and shows that they improve both the power and size properties of the test in applications such as ours.4

2.1 Identifying structural breaks in economic growth

At the outset, we must decide on the minimum “interstitiary period:” the minimum number of years, h, between breaks.5 Imposing a long interstitiary period means that we could be missing true breaks that are fewer than h periods away from each other, or from the beginning or end of the sample period. However, allowing a short interstitiary period implies that some structural break tests may have to be undertaken on data subsamples containing as few as 2 h + 1 observations. In these

---

2 Whether investment booms correlate with upbreaks is less clear, however (Jones and Ollens, 2008). This is a critical issue for “big push” views of development. Massive scaling up of aid flows to finance capital deepening in poor countries has recently been proposed by Sachs and others. Long before, the development literature focused heavily on investment as the vehicle for generating sustained growth in poor countries (see Murphy et al., 1989; Nurkse, 1953; Rosenstein-Rodan, 1943). The finding that investment is not a strong correlate of upbreaks would appear to cast doubt on these views.

4 Of the various alternatives presented in Antoshin et al. (2008), the sequential procedure with heteroskedasticity correction performs best for time series with low-medium autocorrelation and is thus employed here.

5 Given a sample size T, the interstitiary period h will determine the maximum number of breaks, m for each country: m = int(T/h) – 1. For example, if T = 50 and h = 8, then m = int(6.25) – 1 = 5. In practice, we set m = int(T/h) – 2 to avoid occasional anomalies.
circumstances, the size of the test may no longer be reliable, and the power to reject the null hypothesis of no structural break on the sub-sample may be low. Moreover, we hypothesize that breaks at shorter frequencies may have different determinants, and in particular may embody cyclical factors that we are less interested in here. Balancing these factors, we set \( h \) either equal to 8 or to 5.

We next employ an algorithm that sequentially tests for the presence of up to \( m \) breaks in the GDP growth series. The first step is to test for the null hypothesis of zero structural breaks against the alternative of one or more structural breaks (up to the pre-set maximum \( m \)). The location of potential breaks is decided by minimizing the sum of squared residuals between the actual data and the average growth rate before and after the break. Critical values are generated through Monte Carlo simulations using bootstrapped residuals that take into account the properties of the actual time series (that is, sample size and variance). We choose these critical values so as to reject true nulls at the 10 percent level.\(^6\)

Table 1 and Fig. 1 summarize the results from applying these tests to income per capita growth series in 140 countries for which internationally comparable output data are available since at least the 1970s. Our data source is version 6.2 of the Penn World Tables, extended from 2004 to 2006 using the IMF’s World Economic Outlook database.\(^7\) Table 1 shows the number of “upbreaks” and “downbreaks”, at the 10 percent significance level and minimum interstitial periods (\( h \)) of five and eight years, respectively, by region and decade.\(^5\)

With \( h = 5 \), our algorithm identifies a total of 280 breaks—140 “upbreaks” and 140 “downbreaks”, that is, a little more than one upbreak and one downbreak per country on average. This is dramatically higher than the total number of breaks (74) that the standard Bai-Perron algorithm identifies using the same data, \( p \)-value and interstitial period, and is consistent with the findings reported in Antoshin et al. (2008). Upbreaks tend to be most common in the 1950s and 60s, driven by Europe and Latin America, and in the 1990s, driven by Africa. Downbreaks are particularly concentrated in the 1970s. For the high-income countries, the first half of the 1970s stands out; for Latin America, the period between 1978 and 1983; and for Africa, the 1970s and the first half of the 1980s.

Setting \( h = 8 \) substantially reduces the total number of breaks. We find 174 breaks: 78 upbreaks and 96 downbreaks (the standard Bai-Perron approach identifies only 64 breaks in total). The fact that setting \( h = 8 \) leads to 40% fewer breaks shows that the interstitiary period matters, but it does not tell us which approach is better. With \( h = 5 \), we may be picking up some breaks in long-term growth that we might be missing when we require breaks to be at least eight periods apart. However, we may also be picking up abrupt output movements at shorter frequencies that reflect volatility, business cycles, or short-lived commodity price booms or busts. These are more likely to be filtered out by setting \( h = 8 \). On balance, we prefer to focus on the \( h = 8 \) case. For robustness, we also look at the \( h = 5 \) case.

2.2. From structural breaks to growth spells

The period following a growth upbreak can be thought of as a “growth spell” a time period of higher growth than before, ending either with a downbreak or with the end of the sample. However, it is sometimes the case (after periods of very high growth) that high growth continues, albeit at a lower level. In this case, one would not want to say that a growth spell has ended. Conversely, it is sometimes the case that an upbreak follows a period of sharply negative growth, leading to a period in which growth is still negative (or positive but very small). In this case, one would not want to say that a growth spell is underway.

In short, if the objective is to understand the determinants of desirable growth spells, the statistical criteria discussed in the previous section need to be supplemented by an economic criterion. We hence define growth spells as periods of time

- **beginning** with a statistical upbreak followed by a period of at least \( g \) percent average growth; and
- **ending** either with a statistical downbreak followed by a period of less than \( g \) percent average growth (“complete” growth spells) or with the end of the sample (“incomplete” growth spells).

Since growth in our definition means per capita income growth, growth of as low as 2% might be considered a reasonable threshold. We used \( g = 2 \), \( g = 2.5 \), and \( g = 3 \), with similar results, and focus on the \( g = 2 \) case below.\(^5\)

We now characterize the growth spells that result from applying these criteria to the structural breaks summarized in Table 1, using \( g = 2 \), from several angles.\(^10\)

2.3. Duration of spells

Regions do not differ much in terms of the frequency of growth spells. Table 2 presents the number of growth spells by region together with some rudimentary information about the distribution of the length of these spells. There have been a total (both complete and incomplete) of 64 spells for the \( h = 8 \) case and 104 for \( h = 5 \). A little under half of the spells identified at each level correspond to Latin America and Africa, about in line with the fraction of Latin American and African countries in the sample. Hence, in spite of the potential bias against finding growth spells in these countries as a result of their high year-to-year volatility, Latin America and Africa do not, on average, appear very unusual with respect to their ability to get growth going.

Instead, the real problem in these regions seems to be their inability to sustain growth over long periods. Irrespective of which minimum interstitial period and \( p \)-level we choose, the mean length of growth spells is always much shorter—by up to a half—for Latin America and Africa compared to the industrial countries and emerging Asia.

The table also shows an interesting asymmetry between complete and incomplete growth spells for Africa and Latin America. Latin America had a fair number of (albeit short) growth spells in the past, but it has few ongoing growth spells. In contrast, in Africa a large number of countries have been enjoying ongoing growth spells. Most of these were initiated in the mid to late 1990s which is why

---

\(^6\) The sample size is often small when we reject true nulls 10 percent of the time. In the on-line appendix, we provide a full set of results for tests of 25 percent size. Most of the results presented below hold here as well. Several, however, are weaker for the \( h = 5 \) and \( p = 25 \) sample, suggesting to us that this sample may be excessively contaminated by business cycle effects.

\(^7\) We prefer to use the Penn World Tables because it has been the standard data set in the literature, using the IMF’s data only because it is available more recently. On some of the issues with the Penn World Tables for growth data and a comparison with other data sources, see Johnson et al. (2009).

\(^8\) We also ran the algorithm using higher \( p \)-values to give us a chance to detect more breaks (albeit at the expense of more “false positives”) in countries in which the year-to-year volatility of output is high. This increases the total number of breaks identified, but does not substantially affect the distributions of upbreaks and downbreaks across regions and time periods (available in on-line appendix).

\(^9\) It could be argued that, in defining spells, an additional economic criterion might be applied to the size of the growth break, in addition to that imposed on the level of growth before and after the break. For example, each break might be required to represent a change in growth of at least two percentage points. On balance, we believe that such a requirement is inappropriate: if growth is very smooth in a particular country, such that even a small growth break of say one percentage point stands out statistically, then we should try to understand the determinants of this break. However, we have investigated the properties of the set of spells and the duration regressions that results from the application of this additional minimum break size criterion and report briefly on them below.

\(^10\) In the definition above, growth spells are required to begin with a within-sample upbreak. But there are many country cases in which there are no upbreaks in the first 20 years or so because growth started out high. This growth period could reasonably be regarded as a growth spell initiated by an upbreak outside the sample period. It would be useful to extend the output series backwards, or use information about economic history, to roughly “time” the beginnings of these early growth spells.
3.1. Empirical strategy

The growth spell.

the beginning of the growth spell and to policies undertaken during
equivalently, the probability that a spell will continue beyond a spe-

The lack of data leads us to the following approach. We sequen-
tial test the relevance of particular regressors of interest, while in-
cluding some minimal controls. In the empirical section below, we
begin by running regressions of duration on various proxies for exter-
nal shocks, controlling only for per capita income levels. This is ac-
ceptable if external shocks are not correlated with other (e.g.,
institutional and policy) variables that may be correlated with growth
spells. Finding that some of these shocks matter, we then control for
them while sequentially testing first for the relevance of some institu-
tional variables and income distribution, and then for a variety of
health and education related variables, variables related to trade
and competitiveness, and macroeconomic policy. This sequence is
motivated by the idea that external shocks, institutions, and social
heterogeneity may affect the economy both directly and through
policies. Hence, omitting policies in a model that accounts for shocks,
institution and inequality/heterogeneity merely changes the inter-
pretation of the results but does not necessarily misspecify the
model. At the end, we summarize by showing the results of a few par-
simonious regressions that control for all or most of the variables that
were found to matter during the sequential testing process.

The second challenge relates to the distinction between initial con-
ditions at the beginning of the spell and changes in determinants of du-
ration as the spell proceeds, and to how potential reverse causality can
be addressed in that context. In many cases, we will be studying poten-
tial determinants of spell length that changes over the course of the
spell. This indicator needs to be regarded as endogenous in the sense
that its level may depend on whether the country is in a growth spell
or not. At the same time, however, it might be amenable to policy ac-
tions while a spell is ongoing. Hence, it would be desirable to under-
stand not only how initial conditions affect duration, but also how
ongoing changes in particular variables influence the probability that
a spell will end.

3.2. Growth before, during, and after growth spells

In addition to the incidence and duration of growth spells, overall
growth performance will of course depend on growth levels both during
and between spells. Table 3 examines whether there are systematic dif-
fences across regions in this regard, and also looks at growth immedia-
tely before and after growth spells to see whether there is any
suggestion that growth spells begin or end with economic crises.

In general, there are no big differences in growth levels during
spells across regions (the main exception is Latin America, where
growth spells that began in our sample period have tended to be
somewhat less vigorous than in other countries). In contrast, there
are big differences with respect to growth after spells ended. In the
advanced countries and Asia, growth spells have on average ended
with (relatively) “soft landings”—growth rates between −1 and
3%—while African spells have tended to end with deep collapses,
with average growth rates between −3 and −6%. The remaining de-
velling countries occupied an intermediate position, with growth
rates between −1 and −3%.

There are also interesting differences in growth rates before the
onset of growth spells, particularly for spells that are currently incom-
plete. Asian and high-income countries tend to start their spells from
per capita growth rates that are positive or very slightly negative. In
contrast, growth spells in the remaining developing country regions
tend to begin with crises. In these regions, average interstitial rates
prior to the last round of growth spells are between −1 and −5%,
with even lower rates immediately prior to the onset of growth spells.

3. Analyzing duration of growth spells

We would like to relate the expected duration of growth spells—or
equivalently, the probability that a spell will continue beyond a spec-
cific length—to the economic and political conditions prevailing at
the beginning of the growth spell and to policies undertaken during
the growth spell.

3.1. Empirical strategy

We face two main challenges. The first concerns model selection. The
approach of this paper is atheoretical in the sense that we do not base
our modeling priors on a particular theory of why growth is more
sustained in some cases than in others. Instead, our priors are influenced
by a variety of ideas from the existing literature. While these ideas give
us a bit of structure that helps us think about model selection, they also
point to a very wide range of potential determinants of growth duration
and to possible interactions among those determinants.

So far, this does not sound very different from the standard model
selection problem in empirical growth analysis, which has received sig-
nificant attention in recent years (for example, Fernández et al., 2001;
Hendry and Krolzig, 2004; Hoover and Perez, 2004; Sala-i-Martin et
al., 2004). It is greatly complicated in our case, however, by data con-
straints that preclude the application of a general-to-specific modeling
approach. The problem is that in order to analyze growth spells that
began as far back as the 1950s for 119 countries, we require data for
many countries over very long time periods. Few long data series are
available for many countries. Furthermore, the sample of growth spells
that we start out with is small. As a result, running any growth spells re-
gression that includes the main “usual suspects”—to say nothing of a
broader regression that includes a number of other variables or interac-
ctions—will shrink degrees of freedom to uncomfortably low levels.

The lack of data leads us to the following approach. We sequen-
tially test the relevance of particular regressors of interest, while in-
cluding some minimal controls. In the empirical section below, we
begin by running regressions of duration on various proxies for exter-
nal shocks, controlling only for per capita income levels. This is ac-
ceptable if external shocks are not correlated with other (e.g.,
institutional and policy) variables that may be correlated with growth
spells. Finding that some of these shocks matter, we then control for
them while sequentially testing first for the relevance of some institu-
tional variables and income distribution, and then for a variety of
health and education related variables, variables related to trade
and competitiveness, and macroeconomic policy. This sequence is
motivated by the idea that external shocks, institutions, and social
heterogeneity may affect the economy both directly and through
policies. Hence, omitting policies in a model that accounts for shocks,
institution and inequality/heterogeneity merely changes the inter-
pretation of the results but does not necessarily misspecify the
model. At the end, we summarize by showing the results of a few par-
simonious regressions that control for all or most of the variables that
were found to matter during the sequential testing process.

The second challenge relates to the distinction between initial con-
ditions at the beginning of the spell and changes in determinants of du-
ration as the spell proceeds, and to how potential reverse causality can
be addressed in that context. In many cases, we will be studying poten-
tial determinants of spell length that changes over the course of the
spell. This indicator needs to be regarded as endogenous in the sense
that its level may depend on whether the country is in a growth spell
or not. At the same time, however, it might be amenable to policy ac-
tions while a spell is ongoing. Hence, it would be desirable to under-
stand not only how initial conditions affect duration, but also how
ongoing changes in particular variables influence the probability that
a spell will end.

Table 1

Growth breaks by decade and region.

<table>
<thead>
<tr>
<th>Region</th>
<th>Minimum segment = 5</th>
<th>Minimum segment = 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of countries</td>
<td>Total</td>
<td>Average break size</td>
</tr>
<tr>
<td>Total upbreaks</td>
<td>140</td>
<td>137</td>
</tr>
<tr>
<td>Industrial countries</td>
<td>37</td>
<td>27</td>
</tr>
<tr>
<td>Emerging Asia</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>28</td>
<td>33</td>
</tr>
<tr>
<td>Africa and Middle East</td>
<td>53</td>
<td>52</td>
</tr>
<tr>
<td>Total downbreaks</td>
<td>140</td>
<td>150</td>
</tr>
<tr>
<td>Industrial countries</td>
<td>37</td>
<td>38</td>
</tr>
<tr>
<td>Emerging Asia</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>28</td>
<td>34</td>
</tr>
<tr>
<td>Africa and Middle East</td>
<td>53</td>
<td>58</td>
</tr>
</tbody>
</table>

Note: a Includes Japan, Korea, Singapore, Hong Kong SAR, and Taiwan, Province of China.
We seek to address the issue of endogeneity by distinguishing, for most variables, between the initial level of the variable at the beginning of the spell, and changes since the beginning of the spell. Reverse causality is addressed by estimating the relationship between these time-varying variables and expected duration of a spell conditional on its current length (i.e. conditional on being in an ongoing spell). As explained in more detail below, this is achieved through a survival model with time-varying covariates, that can be viewed as roughly analogous to a panel estimation in which the right hand side variables are predetermined (though not strictly exogenous). While this will not eliminate all sources of endogeneity (for example, endogeneity through expectation that the end of a spell is imminent), it should
Table 2
Frequency and duration of growth spells.

<table>
<thead>
<tr>
<th>No. of countries</th>
<th>Minimum length of spell: 5 years</th>
<th></th>
<th>Minimum length of spell: 8 years</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of spells</td>
<td>Mean duration</td>
<td>% of spells lasting at least</td>
<td>No. of spells</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10 years</td>
<td>16 years</td>
</tr>
<tr>
<td>Complete spells</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industrial countries</td>
<td>37</td>
<td>9</td>
<td>12.7</td>
<td>67</td>
</tr>
<tr>
<td>Emerging Asia</td>
<td>22</td>
<td>6</td>
<td>21.3</td>
<td>83</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>18</td>
<td>11</td>
<td>9.9</td>
<td>27</td>
</tr>
<tr>
<td>Africa and Middle East</td>
<td>43</td>
<td>11</td>
<td>5.7</td>
<td>0</td>
</tr>
<tr>
<td>Other developing</td>
<td>20</td>
<td>12</td>
<td>8.6</td>
<td>17</td>
</tr>
<tr>
<td>Incomplete spells</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industrial countries</td>
<td>37</td>
<td>11</td>
<td>25.5</td>
<td>91</td>
</tr>
<tr>
<td>Emerging Asia</td>
<td>22</td>
<td>14</td>
<td>20.4</td>
<td>71</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>18</td>
<td>3</td>
<td>22.7</td>
<td>100</td>
</tr>
<tr>
<td>Africa and Middle East</td>
<td>43</td>
<td>19</td>
<td>14.2</td>
<td>68</td>
</tr>
<tr>
<td>Other developing</td>
<td>20</td>
<td>8</td>
<td>20.1</td>
<td>75</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industrial countries</td>
<td>37</td>
<td>20</td>
<td>19.8</td>
<td>80</td>
</tr>
<tr>
<td>Emerging Asia</td>
<td>22</td>
<td>20</td>
<td>20.7</td>
<td>75</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>18</td>
<td>14</td>
<td>12.6</td>
<td>43</td>
</tr>
<tr>
<td>Africa and Middle East</td>
<td>43</td>
<td>30</td>
<td>11.1</td>
<td>43</td>
</tr>
<tr>
<td>Other developing</td>
<td>20</td>
<td>20</td>
<td>13.2</td>
<td>40</td>
</tr>
</tbody>
</table>

a Includes Japan, Korea, Singapore, Hong Kong SAR, and Taiwan, Province of China.
b Middle East, North Africa, Cyprus, Turkey, and Caribbean countries.

Table 3
Average growth before, during and after growth spells.

<table>
<thead>
<tr>
<th></th>
<th>Minimum length of spell: 5 years</th>
<th></th>
<th>Minimum length of spell: 8 years</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average growth</td>
<td>3 years</td>
<td>Average growth</td>
<td>3 years</td>
</tr>
<tr>
<td></td>
<td>Before</td>
<td>During</td>
<td>After</td>
<td>Before start</td>
</tr>
<tr>
<td>Complete spells</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industrial countries</td>
<td>0.6</td>
<td>7.9</td>
<td>−0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>Emerging Asia</td>
<td>−1.5</td>
<td>7</td>
<td>−1.3</td>
<td>−1.3</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>0.6</td>
<td>4.7</td>
<td>−0.1</td>
<td>0</td>
</tr>
<tr>
<td>Africa and Middle East</td>
<td>−0.9</td>
<td>9.5</td>
<td>−3.2</td>
<td>−3.1</td>
</tr>
<tr>
<td>Other developing</td>
<td>−1.8</td>
<td>6</td>
<td>−2.3</td>
<td>−1.5</td>
</tr>
<tr>
<td>Incomplete spells</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industrial countries</td>
<td>−0.7</td>
<td>5.9</td>
<td>−1.6</td>
<td>0.1</td>
</tr>
<tr>
<td>Emerging Asia</td>
<td>−0.6</td>
<td>4.9</td>
<td>−0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>−2</td>
<td>3.2</td>
<td>−3.9</td>
<td>−1.3</td>
</tr>
<tr>
<td>Africa and Middle East</td>
<td>−5.1</td>
<td>7.5</td>
<td>−7.8</td>
<td>−4.3</td>
</tr>
<tr>
<td>Other developing</td>
<td>−2.9</td>
<td>5.9</td>
<td>−5.4</td>
<td>−2.8</td>
</tr>
</tbody>
</table>

a Includes Japan, Korea, Singapore, Hong Kong SAR, and Taiwan Province of China.
b Middle East, North Africa, Cyprus, Turkey, and Caribbean countries.

3.2. Regression methodology

Let \( t \) denote “analysis time” (time since growth accelerated) and \( T \) duration (the length of a growth spell), a random variable. Thus \( t = 1 \) denotes the first year in a growth spell. \( X(t) \) is a vector of random variables that may influence the probability that a growth spell ends; \( x_t \) is the realization of \( X(t) \) at time \( t \); and \( z \) is a vector of non-time varying variables that may also have an impact on the length of a growth spell. \( z \) could contain realizations of \( X(t) \) before the beginning of a growth spell (i.e. \( x_{t<1} \)) and also variables that have no time dimension at all, e.g. geographical variables. We want to estimate the relationship between the \( X(t) \) and \( z \) variables and \( T \).

Duration is usually modeled by parameterizing the hazard rate—the conditional probability that the spell will end in the next period—and estimating the relevant parameters using maximum likelihood. The most popular approach is to assume a “proportional hazard model”—in effect, an assumption that the time dependence of the hazard, called the “baseline hazard,” is multiplicatively separable from its dependence on other variables \( X(t) \) and \( z \)—and to parameterize it by assuming that the relationship between the hazard and these other variables is linear and that the “baseline hazard” takes a particular functional form:

\[
\lambda(t) = g(X(t), z) \lambda_0(t) = \exp[\beta^T X(t), z] \lambda_0(t) \tag{1}
\]

where \( \lambda_0(t) \) is assumed to obey a specific distribution whose parameters can be estimated along with the coefficient vector \( \beta \).

One potential problem in estimating Eq. (1) arises from the feedback of duration to the covariates \( X \); i.e. the fact that \( X \) might depend on whether or not a spell has ended or is still ongoing. As shown by Woolridge (2002), Eq. (1) can be estimated consistently if we can assume that the hazard at time \( t \) conditional on the covariates at time \( t \) depends only on the lagged realizations of those covariates, i.e. when it depends neither on future realizations of the covariates nor on unobserved covariates.
Intuitively, we make three important assumptions. First, we rule out contemporaneous feedback from the end of a growth spell to the time-varying covariates within the current time period. Second, we must assume that duration is conditionally independent of censoring. This is automatically satisfied in our sample since we have fixed censoring (all growth observations end in 2006). Third, and as usual, we must not omit relevant variables from the regression. Given the data availability constraints discussed at the end of the last section, this is potentially the most serious problem.

The nature of duration dependence itself—that is, the effect of time in spell on the hazard—is of some intrinsic interest. Do spells tend to become more sustainable over time, or is there a sense in which there is a natural spell length? Of course, we would like to say that we have explained such behavior through the evolution of other variables included in the regression, so in a sense significant duration dependence is a sign of modeling failure. At a minimum, though, it is important to know if our other results are sensitive to assumptions about the nature of duration dependence.

We have used as a baseline specification the assumption that \( \lambda(t) \) follows a Weibull distribution, i.e. \( \lambda(t) = \lambda_0(t)^p \). The parameter \( p \), which is estimated, determines whether duration dependence is positive \((p>1)\) or negative. In the robustness section below, we also look at a number of alternative functional forms, including some that allow duration dependence to be non-monotonic. The main results are robust to alternative distributional assumptions.

A final methodological issue is that we treat our dependent variable (the length of the spell) as known, when in fact it is estimated and thus observed with error. This problem, which is a general affliction of this and related literatures, is not easy to solve. Abrevaya and Hausman (1999) and Lancaster (1985) show that hazard estimates using the Weibull distribution are consistent even with measurement error in the dependent variable. However, our setup does not meet the requirements for this result (most notably, it applies only to non-censored data). We have conducted some simulations in an attempt to gauge the direction and magnitude of the bias that may be associated with this measurement error. These exercises suggest that, for our baseline regression and data, there is indeed likely to be some attenuation of the estimated association between the hazard and its correlates—the estimated magnitude of \( \beta \) in Eq. (1)—relative to the true value, due to the measurement error induced by the fact that the growth breaks and spells are estimated rather than directly observed. This result provides some comfort that our results are “conservative” in that, if we find a significant correlation, it would likely be stronger if we could directly observe the growth spells with certainty.

3.3. Results

We start by characterizing the unconditional hazard rate, or the probability that a spell will end after a given number of years, conditional only on the fact that it has already lasted up to that point. We then examine the role of external shocks, then of institutions and variables related to social conflict (income distribution and ethnic heterogeneity), and then of a variety of other policy-related indicators, using some of the previous variables as controls.

3.3.1. Unconditional hazard

Fig. 2. Unconditional hazard. Notes: based on weighted kernel smoother of the empirical hazard rate, for minimum interstitial period of 5 years, using data from 64 spells with minimum length of 8 years.

![Fig. 2](image)

The table shows unconditional hazard for spells defined with a minimum interstitial period of eight years. This hazard is calculated simply by looking at all the spells that have lasted at least a given length and calculating how many end in the next period (then smoothing). This hazard declines with time, implying that spells have already lasted longer have a lower hazard of ending. We will now examine how spell duration is correlated with various determinants of interest. Armed with this analysis, we can return to the question of whether, controlling for the evolution of these covariates, there is any intrinsic time effect on the hazard.

3.3.2. External shocks

We focus on two external shocks: changes in the terms of trade and changes in nominal U.S. interest rates. We include terms of trade shocks (measured as year-to-year percentage changes) and contemporaneous and first lags for U.S. interest rate changes. Table 4 shows the results from running these models on our two samples: interstitial periods of eight and five years with breaks identified at the \( p = 0.1 \) level.

The table shows exponentiated regression coefficients. These can be interpreted as “time ratios”: the factor by which the time to failure is multiplied as a result of increasing the value of the regressor by one unit. For example, a time ratio of 0.9 means that a unit change in the regressor decreases the expected time to failure by 10%. A time ratio of 1 means there is no effect, and a time ratio greater than 1 denotes a “protective” relationship. The \( p \)-values shown adjacent to the time ratios refer to the probability that the true time ratio equals 1. We use the convention that time ratios that are significantly different from 1 at the 5 percent level or less are denoted in bold and time ratios significant at the 10 percent but not 5 percent level in bold and italics. As expected—given Rodrik (1999) and related work—external shocks are associated with a higher risk that growth spells end. For the terms of trade, a time ratio of 1.03–1.04 means that a one percentage point improvement in the terms of trade will increase the expected length of the spell by 3–4%. We also find a very large and significant association of U.S. interest rate changes with duration: depending on the sample, a one percentage point (100 basis points) increase in U.S. rates in the previous year is estimated to reduce the predicted spell length by up to 56%. Qualitatively similar (though less precise) results are obtained if a 0–1 dummy variable for large

---

13 We use the Stata command Strexp with the accelerated-time-to-failure option. To take account of the fact that a downbreak, by construction, cannot happen until five (or eight) years into a spell at the earliest (a consequence of our interstitial period), we created a dummy variable defining the notional start of the spell as the true start year plus four years.

14 These exercises involve: (1) simulating growth data that resemble the growth data we use in our regressions and which includes growth spells whose hazards by construction follow a conditional Weibull distribution such as that estimated in one of our benchmark regressions; (2) estimating hazard regressions on the growth spells in these simulated data, either (a) using the true breaks from the simulated data or (b) the breaks in the simulated data as estimated by our modified Bai–Perron procedure; and (3) comparing the estimates in 2a and 2b. These results are available on request.

15 Appendix Table 1 provides summary statistics for all the regressors.

16 A similar picture obtains for \( h = 5 \), though with a suggestion that the hazard is flat after falling for the first few years.

17 Results for breaks identified at the \( p = 0.25 \) level are available in the online appendix.
hikes in the U.S. federal funds rate is used instead of the continuous variable for U.S. market rates (as in Becker and Mauro, 2006).

3.3.3. Political and economic institutions

There is a well-established link between long-run growth and political institutions—in particular, the extent to which they achieve political accountability and constrain the executive. Whether this link operates via more vigorous growth or more sustained growth spells or both is not obvious. One channel through which weaker institutions could lead to shorter growth spells is by making societies deal less effectively with external shocks (Rodrik, 1999). Another is that poor institutions breed economic and political problems that make countries more crisis-prone and growth more volatile (Acemoglu et al., 2003).

Table 5 shows the relationship between institutions and the length of growth spells in our data. The standard “polity2” measure of democratic institutions, measured on a scale of 0–10 (most autocratic) to 10+ (most democratic) shows a “protective” relationship. A one point improvement in the polity score is associated with an increase in the expected duration of a growth spell by 10–12% (model 1). This holds for both cross-sectional differences in the polity of societies at the beginning of a growth spell and for changes in the polity score within a growth spell. Since the coefficient on these “initial level” and “change within spell” variables are similar, one can collapse them and simply control for the contemporaneous level of the polity variable with the effect that the precision of the estimate is somewhat improved (see model 2, where the increase in predicted duration associated with a one point increase in the polity score is 12%).

Models 3 and 4 in Table 5 show that the strong relationship between the polity variable and the time ratio is driven by both of its main subcomponents, namely, separate assessments of “democracy” and “autocracy”, which are both measured on a 0–10 point scale (“polity” is defined as the democracy score minus the autocracy score). Models 5–7 examine three “concept variables” in the polity database on which both the “democracy” and “autocracy” scores are based to varying degrees, namely, “executive recruitment” (i.e., how governments come to power), scored between 1 and 8, where larger values mean more democratic); “executive constraints” (i.e., the presence of checks and accountability, scored between 1 and 7, where larger values mean more constraints); and “political competition” (e.g., the relevance of parties and civil society organizations, scored between 1 and 10, where higher values mean more competition). The results are all robust and go in the expected direction.

What about the role of narrower “economic institutions”? These might also matter; for example, even countries with autocratic systems might be able to develop institutions that provide a growth-friendly climate by protecting the rights of investors and entrepreneurs, or property rights more generally. The problem is that direct measures of these institutions are not available over long time periods. The longest available series appear to be those compiled by the International Country Risk Guide (ICRG), beginning in 1984. The eighth model in Table 5 includes an ICRG measure that describes economic, as distinct from political, institutions: “investment profile,” which codes contract enforcement, profit repatriation and payment delays on a 0–12 point scale. The results are consistent with a highly “protective” relationship, with a one point increase (improvement) in investor protection associated with an increase in the expected spell length of 36–64%. However, the estimates are imprecise.

3.3.4. Inequality and fractionalization

Do more homogenous societies—either in terms of income distribution or in terms of ethnic of religious composition—have longer growth spells? This may be the case if growth ends as a result of social or political conflict; or if more homogeneous societies (just like societies with better institutions) are more capable of adapting to shocks. Controlling for terms of trade shocks and U.S. interest rate shocks, we examine two measures of heterogeneity: economic, proxied by the Gini coefficient; and social, proxied by a measure of ethnic heterogeneity (Table 6).

The main result is that there is a large and statistically significant association between income inequality and duration. A one percentage point higher Gini is associated with an expected duration of the growth spell that is lower by between 11 and 15%. Since the cross-sectional standard deviation of the Gini in our sample in 2000, for example, is over 10 percentage points, this is a very strong relationship. Note that, unlike democratization, all the action comes from cross-sectional differences in initial levels of the Gini; “within spell” changes are estimated very imprecisely (which is perhaps not surprising, given the high persistence of the Gini over time) and do not have statistically significant relationships. Model 2 shows that the relationship is preserved if one simply includes the contemporaneous Gini into the model, rather than distinguishing between Ginis at time zero and changes in the Gini. This is important because controlling only for the contemporaneous Gini allows us to work with a large sample that includes a number of extra spells for which initial Ginis were not available.

---


20 Year-to-year Gini proxies were obtained by linearly interpolating levels from the WIDER 2a database of worldwide income inequality (June 2005). The potential mis-measurement resulting from this linear interpolation may be another reason why within-spell changes in the Gini appear to have no effect.
Table 6
Duration regressions: inequality and fractionalization.

<table>
<thead>
<tr>
<th>Model</th>
<th>Variable</th>
<th>8-year minimum spell</th>
<th>5-year minimum spell</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Time ratio</td>
<td>p-value</td>
</tr>
<tr>
<td>1</td>
<td>Inequality (Gini coefficient)</td>
<td>Initial level</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Change within spell</td>
<td>1.08</td>
</tr>
<tr>
<td>2</td>
<td>Inequality (Gini coefficient)</td>
<td>Spells/failures</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ethnic fractionalization (Alesina et al., 2003)</td>
<td>Spells/failures</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Ethnic fractionalization (Alesina et al., 2003)</td>
<td>Spells/failures</td>
</tr>
<tr>
<td>4</td>
<td>Inequality (Gini coefficient)</td>
<td>Spells/failures</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Regressions control for terms of trade shocks, US interest rate changes, and initial income.

Table 7
Duration regressions: social and physical indicators.

<table>
<thead>
<tr>
<th>Model</th>
<th>Variable</th>
<th>8-year minimum spell</th>
<th>5-year minimum spell</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Time ratio</td>
<td>p-value</td>
</tr>
<tr>
<td>1</td>
<td>Primary education</td>
<td>Initial level</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Change within spell</td>
<td>9.19</td>
</tr>
<tr>
<td>2</td>
<td>Secondary education</td>
<td>Initial level</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Change within spell</td>
<td>20.43</td>
</tr>
<tr>
<td>3</td>
<td>Adult mortality</td>
<td>Initial level</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Change within spell</td>
<td>0.96</td>
</tr>
<tr>
<td>4</td>
<td>Child mortality</td>
<td>Initial level</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Change within spell</td>
<td>0.78</td>
</tr>
<tr>
<td>5</td>
<td>Telephone mainlines per 100 people</td>
<td>Initial level</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spells/failures</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Regressions control for terms of trade shocks, US interest rate changes, and initial income.

In contrast, measures of ethnic, linguistic, or religious heterogeneity—available as cross-sectional variables only—do not seem to have a robust significant association with the length of growth spells. Model 3 shows results based on one such variable, namely, an ethnic fractionalization measure compiled by Alesina et al. (2003). It is small in magnitude and insignificant. Related measures by Easterly and Levine (1997) and Fearon (2003) have even weaker relationships. Model 4 shows the estimated time ratios for the Alesina et al. measure, while controlling for income distribution. The estimated time ratio is now significantly below unity. When we repeat the regressions using identical subsamples without controlling for inequality we obtain very similar results (not shown), indicating that the differences using identical subsamples without controlling for inequality are driven by the samples and not by the use of income inequality as a control.

Finally, we investigate the relationship between the duration of growth spells and direct measures of violent conflict—such as wars and internal strife—taken from the Uppsala Conflict Data Program/International Peace Research Institute (UCDP/PRIO) armed conflict dataset (Gleditsch et al., 2002). Surprisingly, we do not find strong robust associations.

There are two interpretations for this. First, high violent conflict, at least according to the UCDP/PRIO scoring, is not as obviously correlated with low growth as one might expect. Simple cross-country correlations indicate a negative but not statistically significant correlation between average internal conflict and average growth and a surprising statistically significant positive correlation between external conflict and growth. Second, it is possible that violent conflicts depress growth primarily through their effects after a growth spell has already ended rather than by ending a growth spell.

3.3.5. Social and physical indicators

We next examine indicators related to education, health, and physical infrastructure, while continuing to control for terms of trade shocks, U.S. interest rate changes, and initial income. Table 7 shows that a number of indicators are significant, though unevenly so across samples. The most robustly significant variable is within-spell improvement in primary education, where the estimated coefficients are very high: a one-month improvement raises predicted duration by some 25–70%. Both the initial level of and increases in child mortality reduce the predicted duration of a spell, with significance only in the larger sample, when \( h = 5 \).

Finally, the time ratio associated with an indicator of physical infrastructure development—telephone mainlines per capita, model 5—is higher than unity as expected, with an increase of one telephone per 100 people associated with a 1 percent increase in the expected spell duration, again only significantly so when \( h = 5 \).

3.3.6. Financial development

Many economists argue forcefully that financial development is a critical part of the growth process for many reasons.\(^{21}\) We find some significant associations between various measures of higher financial development and the duration of growth spells, particularly when \( h = 5 \) (Table 8). We find that an increase in the ratio of bank deposits to GDP, or alternatively of private credit/GDP, of one percentage point is associated with 3 to 4 percent longer spells. The level also seems to have a positive correlation with duration. To maximize the sample size and save degrees of freedom, we look just at contemporaneous (lagged) level effects (models 3 and 4) and find similar results.

3.3.7. Globalization

Trade integration and openness have long been linked to growth performance through links such as market size and competition. In contrast, the relationship between financial integration and growth is more controversial and less robust. Table 9 shows the relationship between trade integration and financial integration proxies and the duration of growth spells. We find a significant and large correlation with trade liberalization—measured by the Wacziarg–Welch dummy variable. Roughly speaking (and bearing in mind that we are controlling only for external shocks and initial income at this point), countries that have liberalized trade appear to enjoy spells that are 3–8 times longer. This effect carries over to openness, though it is significant only with \( h = 8 \) and the relationship is weaker.\(^{22}\) The tension between these results may indicate that the Wacziarg–Welch dummy variable might be a proxy for

\(^{21}\) See e.g. Levine (1997) for a review.

\(^{22}\) Following Pritchett (1996), the measure used here adjusts for cross-country differences in size, access to the sea, distance to export markets, and whether or not the country is an energy producer. That is, openness is measured as the residual in a regression of the sum of exports and imports, as a share of GDP, on these structural characteristics. However, results are similar with unadjusted openness.
reforms that go beyond just trade liberalization, as argued in Rodriguez and Rodrik (2000).

The relationship between financial integration on growth—measured as the sum of external assets and liabilities, expressed, like trade openness, as a share of GDP—is weak or imprecisely estimated. However, a disaggregation of the financial integration turns out to be important. Debt liabilities accumulation may be associated with shorter spells, although the coefficient is small and not significant. In contrast, FDI flows are a significant predictor of longer spells, at least in the $h=5$ sample, with an increase in FDI liabilities by 1% of GDP in the recipient country over the course of the spell associated with an increase in the expected duration of a growth spell by 5–11%. This provides support for the idea that the benefits of financial integration depend on the structure of the assets and liabilities that are exchanged, consistent with the findings in Dell’Ariccia et al. (2008).

3.3.8. The current account, competitiveness, and export structure

Contrary to theories in which growth is constrained by access to capital, several recent papers find that foreign financing does not seem to deliver a growth bonus in developing countries.23 This could be related to the capacity of developing countries to absorb capital inflows, to the macroeconomic consequences of capital inflows, which can lead to overvalued exchange rates which undermine growth in the manufacturing sector, or to the notion that low domestic savings is a sign of poor domestic property rights institutions.

The second interpretation—that reliance on foreign financing is bad because it hurts the development of a manufacturing sector—suggests that the structure of exports (or more generally production) might matter for future growth. Hausmann et al. (2006) suggest a measure of the sophistication of goods that an economy could potentially produce—namely, sophistication weighted by a measure of proximity to the current export basket, based on the frequency with which particular pairs of goods are exported by the same country (see Hausmann et al., 2007; Hidalgo et al., 2007). They argue that this measure, which they call “Open Forest,” should capture the ease with which economies can shift to other export baskets of high sophistication and hence high growth promise, for example, in response to adverse shocks. They show that “Open Forest” is indeed inversely related to the length of stagnation periods. By the same token, it might be positively related to the ability of an economy to sustain growth.

Table 8
Duration regressions: financial development*.

<table>
<thead>
<tr>
<th>Model</th>
<th>Variable</th>
<th>8-year minimum spell</th>
<th>5-year minimum spell</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Time ratio</td>
<td>p value</td>
</tr>
<tr>
<td>1</td>
<td>Bank deposits (percent of GDP)</td>
<td>Initial level</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Change within spell</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spells/failures</td>
<td>37/13</td>
</tr>
<tr>
<td>2</td>
<td>Private credit (percent of GDP)b</td>
<td>Initial level</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Change within spell</td>
<td>1.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spells/failures</td>
<td>37/13</td>
</tr>
<tr>
<td>3</td>
<td>Bank deposits (percent of GDP)</td>
<td>Initial level</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spells/failures</td>
<td>53/19</td>
</tr>
<tr>
<td>4</td>
<td>Private credit (percent of GDP)</td>
<td>Initial level</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spells/failures</td>
<td>53/19</td>
</tr>
</tbody>
</table>

* Regressions control for terms of trade shocks, US interest rate changes, and initial income.

b Private credit by deposit money banks and other financial institutions.

Table 9
Duration regressions: globalization*.

<table>
<thead>
<tr>
<th>Model</th>
<th>Variable</th>
<th>8-year minimum spell</th>
<th>5-year minimum spell</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Time ratio</td>
<td>p value</td>
</tr>
<tr>
<td>1</td>
<td>Trade liberalization (Wacziarg–Welch Dummy Variable)</td>
<td>Initial level</td>
<td>2.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Change within spell</td>
<td>6.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spells/failures</td>
<td>38/16</td>
</tr>
<tr>
<td>2</td>
<td>Trade openness (based on PWT data, adjusted for structural characteristics)</td>
<td>Initial level</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Change within spell</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spells/failures</td>
<td>51/15</td>
</tr>
<tr>
<td>3</td>
<td>Financial integration (sum of external assets and liabilitiesb)</td>
<td>Initial level</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Change within spell</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spells/failures</td>
<td>29/7</td>
</tr>
<tr>
<td>4</td>
<td>External debt liabilitiesb</td>
<td>Initial level</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Change within spell</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spells/failures</td>
<td>29/7</td>
</tr>
<tr>
<td>5</td>
<td>FDI liabilities (Lane and Milesi-Ferretti databaseb)</td>
<td>Initial level</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Change within spell</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spells/failures</td>
<td>29/7</td>
</tr>
</tbody>
</table>

* Regressions control for terms of trade shocks, US interest rate changes, and initial income.

b Lane and Milesi-Ferretti (2007).

Table 10 shows how the variables stressed by this literature relate to the length of growth spells. In general, the results are supportive of the view that current account surpluses, competitive exchange rates, and export structure matter for growth duration.

First, increasing the current account surplus during a growth spell seems to raise the chance that growth will be sustained in the $h=8$ sample, where a one-percentage-point-of-GDP increase in the current account deficit during the spell lowers the expected duration by 26%. Similar results obtain for domestic savings. Increases in the degree of overvaluation during a spell are also correlated with shorter spells in the samples with $h=8$. On these samples, each percentage point increase in overvaluation decreases the predicted spell duration by 2%.

That the current account and overvaluation measures clearly correlate with spell duration in the $h=8$ sample but not in the $h=5$ sample is consistent with the idea that the $h=5$ sample includes spells that reflect business-cycle dynamics. For example, increasing current account deficits and overvaluation may be consistent with a demand-led boom, which could endure for perhaps about five years but would not be ultimately sustainable.

The links between export structure and growth spells are generally large and clear. Manufacturing exports and the length of growth spells show a particularly robust relationship. What seems to matter for sustained growth is not so much the share of manufacturing at the beginning of a growth spell but whether or not manufacturing exports rise as a share of total exports during the growth spell. A one percentage point increase in manufacturing exports is associated with an increase in the expected duration of a growth spell of 3–6%.

Measures of actual or potential export sophistication also seem to correlate with the duration of spells. The coefficients have the expected signs in both levels and changes during spell, and with both measures of sophistication. However, the link is particularly strong and significant for in-spell changes in the “Open Forest” measure proposed by Hausmann and his coauthors. The coefficients are

23 See Aizenman et al. (2004), Prasad et al. (2007), and Rajan and Subramanian (2011).
not as easy to interpret as in the other cases because the export sophistication variables are expressed in natural logs of complex indices; however, the standard deviation of the Hausmann et al. (2007) log export sophistication index is just above 0.5, while the standard deviation of the Hausmann, Rodriguez, and Wagner (log) “Open Forest” index is just above 1. Hence, the results suggest that a one standard deviation increase in export structure flexibility as measured by “Open Forest” roughly doubles the predicted duration of the spell in the h = 8 sample.

3.3.9. Macroeconomic stability

We now examine the relationship between duration and two traditional indicators of macroeconomic volatility: inflation and nominal exchange rate depreciation. In addition, we look for a correlation between growth duration and the average growth rate during the spell. Finally, we examine fiscal performance during a spell, measured in terms of debt/GDP ratios.

For inflation and exchange rate depreciation, we use the traditional log transformation multiplied by 100 to make the time ratios easier to interpret. We could have used ln(n) instead, which some authors (Sarel, 1996; Ghosh and Phillips, 1998) have argued is more appropriate to study the relationship between inflation and growth. However, this transformation would not have worked for exchange rate depreciation (negative depreciations, i.e. appreciations, being very common in our sample) and we want to use the same transformation for both inflation and depreciation to make the coefficients comparable; the results for inflation turn out to be insensitive to the choice of transformation in this case.

The first main result is that nominal instability— inflation or depreciation— appears to be negatively correlated with growth duration, particularly in the sample with h = 8. The results for inflation in this sample suggest that a one point increase in 100 ln(1 + π) during the growth spell—which at low inflation rates is approximately equal to a one percentage point increase in the rate of inflation—is associated with a 1 percent decrease in the predicted duration of the spell. For a depreciation in the exchange rates, the relationship is somewhat stronger and is significant in both samples.

One important question is whether the strong results for inflation and exchange rate depreciation are driven by outliers, as has been argued in the context of conventional cross-country growth regressions (Easterly, 2005). To check this, we drop all observations from the sample in which either current or initial inflation/depreciation exceeded 50% per annum (this means dropping all observations in spells where inflation/depreciation at the beginning of a growth spell exceeded 50%, even if contemporaneous inflation/depreciation was lower). The results (model 3 in Table 9) show that the basic result remains. At the same time, initial inflation rates now have a

---

### Table 10
Regressions: current account competitiveness and export structure.

<table>
<thead>
<tr>
<th>Model</th>
<th>Variable</th>
<th>8-year minimum spell</th>
<th>8-year maximum spell</th>
<th>5-year minimum spell</th>
<th>5-year maximum spell</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Time ratio</td>
<td>p value</td>
<td>Time ratio</td>
<td>p value</td>
</tr>
<tr>
<td>1</td>
<td>Current account balance (percent of GDP, WDI and IFS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Initial level</td>
<td>1.31/0.18</td>
<td>1.02/0.85</td>
<td>1.26/0.09</td>
<td>1.05/0.60</td>
</tr>
<tr>
<td></td>
<td>Spells/failsures</td>
<td>24/6</td>
<td>29/11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Domestic savings (percent of GDP, WDI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Initial level</td>
<td>1.10/0.06</td>
<td>1.03/0.24</td>
<td>1.08/0.13</td>
<td>1.02/0.42</td>
</tr>
<tr>
<td></td>
<td>Spells/failsures</td>
<td>35/11</td>
<td>55/26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Overvaluation (residual of cross-sectional regressions of price levels on PPP GDP per capita)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Initial level</td>
<td>0.99/0.23</td>
<td>1.00/0.61</td>
<td>0.98/0.02</td>
<td>1.00/0.75</td>
</tr>
<tr>
<td></td>
<td>Change within spell</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spells/failsures</td>
<td>52/19</td>
<td>83/42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Manufacturing exports/total exports (percent, WDI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Initial level</td>
<td>1.00/0.87</td>
<td>1.02/0.12</td>
<td>1.06/0.01</td>
<td>1.03/0.03</td>
</tr>
<tr>
<td></td>
<td>Change within spell</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spells/failsures</td>
<td>29/13</td>
<td>43/22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Sophistication of exports (Hausmann et al., 2007)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Initial level</td>
<td>1.82/0.72</td>
<td>5.11/0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Change within spell</td>
<td>7.67/0.19</td>
<td>7.59/0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spells/failsures</td>
<td>39/12</td>
<td>60/31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>“Open Forest” (Hausmann et al., 2006)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Initial level</td>
<td>1.03/0.95</td>
<td>1.54/0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Change within spell</td>
<td>2.04/0.07</td>
<td>1.86/0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spells/failsures</td>
<td>39/12</td>
<td>60/31</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Note:** Regressions control for terms of trade shocks, US interest rate changes, and initial income.

---

### Table 11
Duration regressions: macroeconomic volatility.

<table>
<thead>
<tr>
<th>Model</th>
<th>Variable</th>
<th>8-year minimum spell</th>
<th>8-year maximum spell</th>
<th>5-year minimum spell</th>
<th>5-year maximum spell</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Time ratio</td>
<td>p value</td>
<td>Time ratio</td>
<td>p value</td>
</tr>
<tr>
<td>1</td>
<td>Log (1 + inflation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Initial level</td>
<td>1.00/0.94</td>
<td>1.01/0.58</td>
<td>0.97/0.60</td>
<td>0.99/0.02</td>
</tr>
<tr>
<td></td>
<td>Spells/failsures</td>
<td>54/19</td>
<td>82/43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Log (1 + depreciation in the parallel exchange rate)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Initial level</td>
<td>0.94/0.02</td>
<td>0.99/0.12</td>
<td>0.96/0.06</td>
<td>0.99/0.10</td>
</tr>
<tr>
<td></td>
<td>Change within spell</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spells/failsures</td>
<td>23/9</td>
<td>34/18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Log (1 + moderate inflation)#{2}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Initial level</td>
<td>0.96/0.45</td>
<td>0.91/0.00</td>
<td>0.98/0.64</td>
<td>0.94/0.02</td>
</tr>
<tr>
<td></td>
<td>Spells/failsures</td>
<td></td>
<td></td>
<td>49/19</td>
<td>76/42</td>
</tr>
<tr>
<td>4</td>
<td>Average growth within spell</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spells/failsures</td>
<td>57/19</td>
<td>88/46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Log (1 + inflation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Initial level</td>
<td>1.01/0.84</td>
<td>1.00/0.73</td>
<td>0.98/0.60</td>
<td>0.99/0.01</td>
</tr>
<tr>
<td></td>
<td>Change within spell</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spells/failsures</td>
<td></td>
<td></td>
<td>50/19</td>
<td>82/43</td>
</tr>
<tr>
<td>6</td>
<td>Log (1 + depreciation in the parallel exchange rate)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Initial level</td>
<td>0.98/0.36</td>
<td>0.99/0.06</td>
<td>0.97/0.04</td>
<td>0.98/0.01</td>
</tr>
<tr>
<td></td>
<td>Change within spell</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spells/failsures</td>
<td></td>
<td></td>
<td>50/19</td>
<td>82/43</td>
</tr>
<tr>
<td>7</td>
<td>Debt/GDP change within spell</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spells/failsures</td>
<td></td>
<td></td>
<td>33/8</td>
<td>44/18</td>
</tr>
</tbody>
</table>

---

**Note:** Regressions control for terms of trade shocks, US interest rate changes, and initial income.

---

**Observations with inflation in excess of 50% per annum replaced by missing values.
is among the most robustly significant and economic signiﬁcance of the individual results. Most variables, however, retain their statistical significance at least in one sample. Among the statistically significant variables, the time ratio estimates are for the most part similar to the coefﬁcients examined before, though they tend to be less stable across subsamples. The external shocks variables tend to lose signiﬁcance at least in one sample. The question is whether or not these variables have independent power to predict longer spell duration. It is not obvious that this variable belongs in the spell duration by 2% (model 7). This relationship is less robust absent Table 11 that faster growth is less sustainable. When growth is one percentage point higher, predicted spell duration falls by about 15–36% in all samples and with or without controls for rates of inﬂation or depreciation.

Our third important result is that rapid growth of public debt/GDP during a spell is correlated with shorter duration, at least in the h = 5 sample. The relationship is large economically—a growth of the debt/GDP ratio by 1 percentage point during a spell reduces predicted spell duration by 2% (model 7). This relationship is less robust absent controls for inﬂation and average growth (not shown), perhaps not surprisingly in view of the mechanisms that may link these variables, but which also suggests that, insofar as the relationship is causal, the increase in debt is working through mechanisms other than the effects on aggregate demand or inﬂation.

Of course, it is plausible that many of the correlates of spell duration that we have identiﬁed are themselves correlated with each other. In the next section, we attempt, within the constraints of scarce data, to disentangle the independent contributions of our various regressors.

3.3.10. A summary view

Having concluded our tour of the main covariates that can be usefully analyzed with our data, it is important to see whether the relationships hold up if they are jointly included in the model. Many of the variables that we have identiﬁed as potential correlates of growth spell duration—for example, more equal income distributions and better political institutions—are correlated. The question is whether or not these variables have independent power to predict longer growth spells.

Data scarcity limits the extent to which we can examine the covariates of growth spells jointly. However, it is possible to include at least some variables from each of the groups examined in a way that maintains a reasonable sample size. The results are shown in Table 12 for two versions of the model: one which uses “Overvaluation” among the variable from the competitiveness/export structure group, and the other that uses Hausmann’s open “Open Forest.” In both versions, we also control for terms of trade shocks, U.S. interest rates, and initial income per capita. As in the earlier regressions, a Weibull distribution is assumed.

As expected, the joint inclusion of many variables weakens some of the individual results. Most variables, however, retain their statistical and economic signiﬁcance at least in one sample. Among the statistically signiﬁcant variables, the time ratio estimates are for the most part similar to the coefﬁcients examined before, though they tend to be less stable across subsamples. The external shocks variables tend to lose signiﬁcance, though the signs remain consistent with Table 4. Income inequality is among the most robust predictors of duration. Autocracy (or equivalently polity2, not shown) is also important, depending on the sample and other covariates. Several external variables retain their economic and statistical signiﬁcance in at least one model. Export sophistication as measured by “Open Forest” is among the most robustly signiﬁcant variables. The ratio of manufacturing in total exports turned out not to be signiﬁcant in

Table 12

<table>
<thead>
<tr>
<th>Model</th>
<th>Variable</th>
<th>8-year minimum spell</th>
<th>5-year minimum spell</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time</td>
<td>p value</td>
<td>Time</td>
</tr>
<tr>
<td>1</td>
<td>Log (1 + inflation) (percent, change within spell)</td>
<td>0.99</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>Inequality (Gini coefﬁcient)</td>
<td>0.86</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Autocracy (Polity IV database)</td>
<td>0.70</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>Trade liberalization (Wacziarg-Welch dummy variable)</td>
<td>1.78</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>Overvaluation</td>
<td>0.99</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>External debt liabilities</td>
<td>1.00</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>(Lane and Milesi-Ferretti, 2007)</td>
<td>1.04</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>FDI liabilities (Lane and Milesi-Ferretti, 2007)</td>
<td>1.04</td>
<td>0.09</td>
</tr>
<tr>
<td>2</td>
<td>Log (1 + inflation) (percent, change within spell)</td>
<td>1.00</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Inequality (Gini coefﬁcient)</td>
<td>0.90</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Autocracy (Polity IV database)</td>
<td>0.83</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Trade liberalization (Wacziarg-Welch dummy variable)</td>
<td>1.11</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Open Forest (Hausmann et al., 2006)</td>
<td>2.50</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>External debt liabilities</td>
<td>1.00</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>(Lane and Milesi-Ferretti, 2007)</td>
<td>1.04</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Average growth within spell</td>
<td>0.82</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Terms of trade growth</td>
<td>1.01</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>First lag (US interest rate change)</td>
<td>0.95</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>Initial level of income per capita</td>
<td>0.73</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>FDI liabilities (Lane and Milesi-Ferretti, 2007)</td>
<td>6.04</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>FDI liabilities (Lane and Milesi-Ferretti, 2007)</td>
<td>26/10</td>
<td>1/12</td>
</tr>
</tbody>
</table>

P is the shape parameter in the Weibull distribution. The p-value is for the null hypothesis that p = 1 (implying that the hazard is duration independent).
we have examined, do growth spells have a natural dynamic pattern? This could be interpreted either as intrinsic—how the hazard of growth ending would evolve over time if all covariates were kept constant—or as reflecting time-varying omitted variables (for example, the hazard might rise over the life of a spell as political pressures that tend to bring about an end to the spell build up). In the Weibull specification assumed throughout, the baseline duration dependence—~λ(0) in Eq. (1)—is monotonic and depends on the estimated parameter p, growing with time for p > 1, falling for p < 1, or remains constant if p = 1. We find that in most specifications the estimated value of p is greater than 1, sometimes statistically significantly so (as in model 2 in Table 12), suggesting that the baseline hazard rises with time, controlling for the evolution of the covariates.

3.3.11. Goodness of fit
Do the regressions explain a large share of the variation in growth spell duration? In general, this is a hard question to answer due to the nature of the duration specification with censored observations and time-varying determinants of duration. For example, we would like to calculate the expected spell duration and compare it to the actual. However, unless we assume strict exogeneity, the end of the spell itself will likely change the trajectory of the determinants of duration. Thus, we cannot integrate through all (future) time to calculate expected spell duration. Here, we examine various indicators of goodness-of-fit.

First, we examine the overall statistical performance of our summary model from Table 12 against a series of nested models that have only the most basic covariates or indeed none at all (Table 13). The likelihood-ratio tests imply that the full model fits the data significantly better. As indicated by the high values for the associated chi-squared statistics, these tests reject the full set of zero restrictions implicitly in the smaller models at extremely small p-values (less than 0.01). Along the same lines, we can get a rough sense of the explanatory power of the model by examining a likelihood-based pseudo-R² likelihood ratio statistic measuring goodness-of-fit (Woolridge, 2002). Overall, the covariates seem to have substantial explanatory power in this sense.

Second, we can give a more informal feel for the economic significance of our results. Fig. 3 relates the predicted hazard to the actual incidence of spell failures. To create this figure, we divide all the observations into five equal-sized groups, with the first group having the lowest predicted hazard and the fifth group having highest predicted hazard. For each group, we calculate the average predicted hazard, based on model 2 of Table 12. We can also measure the fraction of observations in each group that represents actual ends of spells. The scatter of the second number against the first represents a picture of “goodness-of-fit.” We also carry out a similar exercise for a restricted model with the same Weibull hazard model but using only per capita income and average growth as predictive variables. We see this figure as suggesting that predicted hazards vary over a wide range—much more so for the summary model—and are reasonably good predictors of the actual frequency of spell failures. This demonstrates that the summary model has substantial explanatory power, much more so than the restricted model.

4. Robustness
In the previous section, we examined the robustness of the results with respect to variations in how exactly breaks in growth, and hence growth spells, were defined. We also looked at how different combinations of potential predictive variables affected the results. This section takes the robustness analysis one step further, examining (1) alternative distributional assumptions about the duration of growth spells, and (2) omitted variables and unobserved heterogeneity.

4.1. Alternative distributional assumptions
The results presented until now have assumed the Weibull distribution for the hazard model. The models presented in the previous section can be re-estimated using a variety of distributional forms. This gives a sense of the robustness of the coefficients of the model to alternative distributional assumptions. It also allows us to examine the robustness of the increasing baseline hazard we observed in the previous section, for example whether allowing for a non-monotonic baseline hazard yields different results. A simple way of generalizing the Weibull distribution is to allow the duration parameter p to be itself a linear function of time as in p = a + bt. This supports a hazard that initially grows with time (p > 1) then falls (p < 1) or vice versa. Similarly, other distributional forms, such as the log-normal, log-logistic, or gamma, will allow hazards that first grow and then fall.

Table 14 shows the results. The columns labeled “Weibull” merely repeat the results for model 2 of Table 12 for reference. The remaining columns estimate the same model using alternative distributions. Unfortunately, the Weibull with an ancillary time parameter could only be estimated for the samples with five-year minimum spell (the estimation would not converge for the other samples using this distribution).

![Fig. 3. Goodness-of-fit. Notes: each data point represents one-fifth of the observations in the sample, ordered by the average predicted hazard. For the summary model, three such groups have zero actual and essentially zero predicted hazard, so they overlap in the figure.](image-url)
The parameter estimates for the log-normal distribution suggest rising baseline hazards with time for the \( t = 5 \) sample and rising then roughly flat baseline hazards for the \( t = 8 \) sample, though with considerable uncertainty about the exact shape. Overall, there is some underwhelming evidence for a rising baseline hazard.

### 4.2. Unobserved heterogeneity and robustness to omitted variables

One potential source of bias in the coefficients estimated in the summary models in Table 12 is an incorrectly specified set of economic covariates. In a strict sense, we know that this must be the case; clearly, it would be a remarkable coincidence if either of the two summary models was exactly right. The best we can hope for is that the specification error that is necessarily present does not greatly affect the coefficient estimates of the variables on which we focus.

But how can this be tested? One approach that is easy to implement (but does not set the bar very high, as we shall see) is to assume that the models are right except for the presence of unobserved heterogeneity (also referred to as “frailty”) in the form of a multiplicative stochastic term \( \upsilon_i \) which is added to the proportional hazard model presented in this section:

\[
\lambda(t) = \upsilon_i \exp[\lambda_0(X(t), z)]
\]

\( \upsilon_i \) is assumed to follow a specific distribution (such as the gamma or inverse Gaussian). Depending on what the index \( i \) pertains to, \( \upsilon_i \) can either be interpreted as a random term modifying the hazard conditional on \( t \) and \( X(t) \) for each observation or as specific to sets of observations (“shared frailty”), for example, across all observations pertaining to a specific country. In this case, \( \upsilon_i \) is analogous to a random effects term in a panel regression context.

Table 15 shows the effect of adding a “frailty” term to the log-gistic version of summary model 2. As can be seen from the table, this barely has an impact on the regression coefficients and significance levels regardless of whether frailty is modeled at the observation level or at the country level.

One potential problem with the frailty regressions in Table 15 is that they assume that the frailty term is uncorrelated with the economic covariates in our model (see Woolridge, 2002). To the extent that unobserved heterogeneity arises because a relevant economic

---

Table 14 shows that the coefficient estimates for the economic covariates are not very sensitive to the assumed distributional form and that their statistical significance often (though not always) is similar across distributions.

With respect to duration dependence itself, as we saw before, when monotonicity is imposed in the Weibull specification, the estimations suggest that the hazard is a rising function of time, with a null of constant hazard (\( p = 1 \)) rejected in model 2 of Table 12 but not in model 1. What happens when monotonicity is not imposed? For model 2, the ancillary time parameter itself is not significant in the augmented Weibull specification, suggesting that we cannot reject monotonicity and thus that the baseline Weibull may be adequate. (The parameter point estimates are consistent with a rising baseline hazard.) The parameter estimates for the log-normal and log-log distributions suggest rising baseline hazards with time for the \( h = 5 \) sample and rising then roughly flat baseline hazards for the \( h = 8 \) sample, though with considerable uncertainty about the exact shape. Overall, there is some underwhelming evidence for a rising baseline hazard.

### Table 14

<table>
<thead>
<tr>
<th>Variable</th>
<th>8-year minimum spell</th>
<th>5-year minimum spell</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Weibull</td>
<td>Log–logistic</td>
</tr>
<tr>
<td>Log ((1 + \text{inflation}))</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>0.83</td>
</tr>
<tr>
<td>Inequality</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Autocracy score</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>Trade liberalization</td>
<td>1.11</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>0.71</td>
<td>0.89</td>
</tr>
<tr>
<td>Open Forest</td>
<td>2.50</td>
<td>2.69</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>External debt liabilities</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>0.09</td>
<td>0.04</td>
</tr>
<tr>
<td>Terms of trade change</td>
<td>1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>U.S. interest rate change</td>
<td>0.82</td>
<td>0.84</td>
</tr>
<tr>
<td>Initial income per capita</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>Constant</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>0.34</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>0.73</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>Ancillary parameter (1))\footnote{a}</td>
<td>6.04</td>
<td>0.14</td>
</tr>
<tr>
<td>Ancillary parameter\footnote{b}</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>p value\footnote{c}</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>Spells/failures</td>
<td>26/10</td>
<td>33/16</td>
</tr>
</tbody>
</table>

\footnote{a}{Reestimation of model 2 from Table 12. Time ratios and \( p \)-values shown, except for Cox, where hazard ratio is presented.}

\footnote{b}{Refers to time coefficient \( b \) of function \( \ln(p) = a + bt \) for Weibull with ancillary time parameter; to shape parameter \( p \) for standard Weibull; to scale parameter \( \gamma \) for log–logistic, and to standard deviation \( \sigma \) for lognormal.}

\footnote{c}{Refers to test that time coefficient \( b = 0 \) in case of Weibull with time parameter; that shape parameter \( p = 1 \) in case of standard Weibull; that \( \gamma = 1 \) in case of log–logistic; and that standard deviation \( \sigma = 1 \) in case of lognormal.}

\footnote{d}{Refers to intercept \( a \) of function \( \ln(p) = a + \delta t \) for Weibull with ancillary time parameter; \( p \)-value refers to test that \( a = 0 \).}

---

We used the log–logistic distribution because attempts to estimate frailty models using the baseline Weibull distribution failed (no convergence). In addition, we tried frailty based on a lognormal distribution with similar results as in Table 15.
covariate has been omitted from the model, this assumption is likely to be violated. An alternative approach to exploring the robustness of the coefficient estimates that is not susceptible to this problem is to add additional variables to the X-matrix of covariates that may proxy for potential omitted variables. In light of the striking differences in the duration of growth spells across regions (see Section 3), good candidates for such variables are regional dummies. Alternatively, one can examine the idea that the sample is heterogeneous by postulating a handful of economic and political characteristics that predict sustained growth: more equal income distribution, democratic institutions, openness to trade and foreign direct investment, and more flexible to shocks (Hausmann et al., 2006; Rodrik, 1999). The results are also consistent with the notion that export orientation may help growth by building constituencies that favor better institutions (Johnson et al., 2006, 2010; Rajan and Zingales, 2006); and the idea that current export or production structures matter for future growth because they favor innovation and allow economies to react more flexibly to shocks (Hausmann et al., 2006, 2007). Exploring, differentiating, and testing these channels remains a challenge for future work.

5. Conclusion

This paper builds on the emerging literature on growth transitions by moving the object of inquiry to the duration of growth spells. Using an extension of Bai and Perron’s (1998, 2003) approach to testing for multiple structural breaks, we identify a rich set of structural breaks in economic growth paths around the world and use these to define “growth spells.” We then employed survival analysis to explore the role of a large number of economic factors that might influence the length of growth spells.

The paper identifies a handful of economic and political characteristics that predict sustained growth: more equal income distribution, democratic institutions, openness to trade and foreign direct investment, and more flexible to shocks. The results are also consistent with the notion that export orientation may help growth by building constituencies that favor better institutions (Johnson et al., 2006, 2010; Rajan and Zingales, 2006); and the idea that current export or production structures matter for future growth because they favor innovation and allow economies to react more flexibly to shocks (Hausmann et al., 2006, 2007). Exploring, differentiating, and testing these channels remains a challenge for future work.

### Table 15
Frailty regressions.

<table>
<thead>
<tr>
<th>Variable</th>
<th>8 year minimum spell</th>
<th>5 years minimum spell</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log-logistic</td>
<td>Log-logistic with frailty</td>
</tr>
<tr>
<td>Log (1 + inflation)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Inequality</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>Autocracy score</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>Trade liberalization</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>Open Forest</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>External debt liabilities</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>FDI liabilities</td>
<td>1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>Avg. growth within spell</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Terms of trade change</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>U.S. interest rate change</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>Initial income per capita</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>Constant</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Scale parameter g</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>p value</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Spells/failures</td>
<td>26/10</td>
<td>33/16</td>
</tr>
</tbody>
</table>

* Refers to test that $\gamma = 1$. 

The paper identiﬁes a handful of economic and political characteristics that predict sustained growth: more equal income distribution, democratic institutions, openness to trade and foreign direct investment, and an export or production structure that favors relatively sophisticated exports. We also ﬁnd that stable macroeconomic environments, with lower inﬂation rates and fewer instances of high depreciation and slower accumulation of debt predict longer growth spells.

The associations and regularities identiﬁed in this paper seem consistent with several themes that have been prominent in the literature on economic development in the last 20 years. These include the view that less equal and cohesive societies experience lower and more volatile growth, perhaps because social conﬂict breeds populist policies (Sachs, 1989) or because they have weaker institutions and a reduced capacity for managing external shocks (Easterly et al., 2006; Rodrik, 1999). The results are also consistent with the notion that export orientation may help growth by building constituencies that favor better institutions (Johnson et al., 2006, 2010; Rajan and Zingales, 2006); and the idea that current export or production structures matter for future growth because they favor innovation and allow economies to react more ﬂexibly to shocks (Hausmann et al., 2006, 2007). Exploring, differentiating, and testing these channels remains a challenge for future work.
<table>
<thead>
<tr>
<th>Model</th>
<th>Variable</th>
<th>8-year minimum spell</th>
<th>5-year minimum spell</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Time ratio</td>
<td>p value</td>
</tr>
<tr>
<td>Model 2</td>
<td>Log (1 + inflation) (percent, change within spell)</td>
<td>1.00</td>
<td>0.95</td>
</tr>
<tr>
<td>(reproduced from Table 12)</td>
<td>Inequality (Gini Coefficient)</td>
<td>0.90</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Autocracy score (scale 0 to 10)</td>
<td>0.93</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Trade liberalization (0–1 dummy)</td>
<td>1.11</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Open Forest</td>
<td>2.50</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>External debt liabilities (percent of GDP)</td>
<td>1.00</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>FDI liabilities (percent of GDP)</td>
<td>1.04</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Average growth within spell (percent)</td>
<td>0.82</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Terms of trade change (percent)</td>
<td>1.01</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>U.S. interest rate change (points)</td>
<td>0.95</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>Initial income per capita (in thousands)</td>
<td>0.73</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Spells/failures</td>
<td>26/10</td>
<td>0.07</td>
</tr>
<tr>
<td>Model 2 with regional dummies</td>
<td>Log (1 + inflation) (percent, change within spell)</td>
<td>1.01</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Inequality (Gini Coefficient)</td>
<td>0.93</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Autocracy score (scale 0 to 10)</td>
<td>0.95</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>Trade liberalization (0–1 dummy)</td>
<td>1.37</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>Open Forest</td>
<td>2.62</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>External debt liabilities (percent of GDP)</td>
<td>1.00</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>FDI liabilities (percent of GDP)</td>
<td>1.03</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Average growth within spell (percent)</td>
<td>0.70</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Terms of trade change (percent)</td>
<td>1.02</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>U.S. interest rate change (points)</td>
<td>0.99</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>Initial income per capita (in thousands)</td>
<td>0.78</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Latin America (dummy)</td>
<td>0.58</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Sub-Saharan Africa (dummy)</td>
<td>1.60</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>Spells/failures</td>
<td>26/10</td>
<td>0.76</td>
</tr>
<tr>
<td>Model 2 w/o Asian observations</td>
<td>Log (1 + inflation) (percent, change within spell)</td>
<td>1.00</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>Inequality (Gini Coefficient)</td>
<td>0.92</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Autocracy score (scale 0 to 10)</td>
<td>0.96</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>Trade liberalization (0–1 dummy)</td>
<td>0.91</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Open Forest</td>
<td>2.55</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>External debt liabilities (percent of GDP)</td>
<td>1.00</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>FDI liabilities (percent of GDP)</td>
<td>1.04</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Average growth within spell (percent)</td>
<td>0.87</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Terms of trade change (percent)</td>
<td>1.02</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>U.S. interest rate change (points)</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Initial income per capita (in thousands)</td>
<td>0.73</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Spells/failures</td>
<td>19/9</td>
<td>25/13</td>
</tr>
</tbody>
</table>

* Re regressions control for terms of trade shocks, US interest rate changes, and initial income.
Appendix A

Appendix Table 1 Summary statistics for regressors.

<table>
<thead>
<tr>
<th>Description</th>
<th>Source</th>
<th>Unit of measurement</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terms of trade</td>
<td>WEO</td>
<td>Ratio of export prices to import prices (2000 = 100)</td>
<td>603</td>
<td>0.48</td>
<td>13.20</td>
</tr>
<tr>
<td>US interest rate</td>
<td>FED 3 month Treasury bill</td>
<td>Percent change</td>
<td>635</td>
<td>−0.02</td>
<td>1.52</td>
</tr>
<tr>
<td>Initial level of income</td>
<td>PWT 6.2</td>
<td>USD per capita</td>
<td>720</td>
<td>3.23</td>
<td>4.47</td>
</tr>
<tr>
<td>Polity 2</td>
<td>Polity IV</td>
<td>Scale from −10 (autocratic) to 10 (democratic)</td>
<td>560</td>
<td>1.22</td>
<td>6.87</td>
</tr>
<tr>
<td>Autocracy</td>
<td>Polity IV</td>
<td>Scale from 0 to 10 (most autocratic)</td>
<td>560</td>
<td>2.97</td>
<td>3.18</td>
</tr>
<tr>
<td>Democracy</td>
<td>Polity IV</td>
<td>Scale from 0 to 10 (most democratic)</td>
<td>560</td>
<td>4.20</td>
<td>3.89</td>
</tr>
<tr>
<td>Executive recruitment</td>
<td>Polity IV</td>
<td>Scale from 1 (ruling by inheritance) to 8 (democratic elections)</td>
<td>560</td>
<td>5.77</td>
<td>2.26</td>
</tr>
<tr>
<td>Executive constraints</td>
<td>Polity IV</td>
<td>Scale from 1 (unlimited executive authority) to 7 (executive parity of subordination)</td>
<td>560</td>
<td>4.22</td>
<td>2.14</td>
</tr>
<tr>
<td>Political competition</td>
<td>Polity IV</td>
<td>Scale from 1 (repressed competition) to 10 (open political participation)</td>
<td>560</td>
<td>5.53</td>
<td>3.29</td>
</tr>
<tr>
<td>Investment profile</td>
<td>ICRG</td>
<td>Scaled from 2 (high risk) to 12 (low risk)</td>
<td>398</td>
<td>7.52</td>
<td>2.22</td>
</tr>
<tr>
<td>Income inequality</td>
<td>Wider 2.a</td>
<td>Scaled from 2 (high risk) to 12 (low risk)</td>
<td>411</td>
<td>38.47</td>
<td>7.36</td>
</tr>
<tr>
<td>Ethnic fractionalization</td>
<td>Alesina et al. (2003)</td>
<td>One minus the sum of ethnic group shares</td>
<td>692</td>
<td>44.16</td>
<td>25.56</td>
</tr>
<tr>
<td>Primary education</td>
<td>Barro and Lee, 2001</td>
<td>Average years of primary schooling in the total population over age 25</td>
<td>395</td>
<td>2.10</td>
<td>1.21</td>
</tr>
<tr>
<td>Secondary education</td>
<td>Barro and Lee, 2001</td>
<td>Average years of secondary schooling in the total population over age 25</td>
<td>395</td>
<td>0.63</td>
<td>0.54</td>
</tr>
<tr>
<td>Adult mortality</td>
<td>WDI</td>
<td>Deaths per 100</td>
<td>502</td>
<td>32.78</td>
<td>11.81</td>
</tr>
<tr>
<td>Child mortality</td>
<td>WDI</td>
<td>Deaths per 100</td>
<td>515</td>
<td>8.56</td>
<td>4.49</td>
</tr>
<tr>
<td>Telephone infrastructure</td>
<td>WDI</td>
<td>Mainlines per 100 people</td>
<td>448</td>
<td>107.98</td>
<td>154.43</td>
</tr>
<tr>
<td>Bank deposits</td>
<td>Beck et al., 2000</td>
<td>Ratio of deposit money bank assets to GDP, in percent</td>
<td>302</td>
<td>29.83</td>
<td>34.48</td>
</tr>
<tr>
<td>Private credit</td>
<td>Beck et al., 2000</td>
<td>Ratio of private credit by deposit money banks to GDP, in percent</td>
<td>302</td>
<td>26.14</td>
<td>28.95</td>
</tr>
<tr>
<td>Trade openness</td>
<td>PWT 6.2</td>
<td>Exports plus Imports divided by GDP, percent</td>
<td>570</td>
<td>6.69</td>
<td>44.27</td>
</tr>
<tr>
<td>Financial integration</td>
<td>Lane and Milesi-Ferretti, 2007</td>
<td>Sum of total foreign assets and liabilities (net of the value of derivatives) as a ratio of GDP.</td>
<td>242</td>
<td>134.41</td>
<td>136.38</td>
</tr>
<tr>
<td>External debt liabilities</td>
<td>Lane and Milesi-Ferretti, 2007</td>
<td>External debt liabilities from IMF’s (WEO) and World Bank’s Global Development Finance database</td>
<td>242</td>
<td>53.81</td>
<td>50.61</td>
</tr>
<tr>
<td>FDI liabilities</td>
<td>Lane and Milesi-Ferretti, 2007</td>
<td>Data on foreign assets and liabilities of banks and other banking institutions reported by IFS</td>
<td>242</td>
<td>16.35</td>
<td>21.52</td>
</tr>
<tr>
<td>Sophistication of exports</td>
<td>Hausmann et al., 2007</td>
<td>Weighted sum of the productivity levels associated to each exported good</td>
<td>414</td>
<td>8.37</td>
<td>0.38</td>
</tr>
<tr>
<td>Current account balance</td>
<td>WDI and IFS</td>
<td>Percent of GDP</td>
<td>210</td>
<td>−3.37</td>
<td>6.27</td>
</tr>
<tr>
<td>Domestic savings</td>
<td>WDI</td>
<td>Percent of GDP</td>
<td>350</td>
<td>15.37</td>
<td>12.56</td>
</tr>
<tr>
<td>Manufacturing exports/total exports</td>
<td>WDI</td>
<td>Ratio of manufacturing goods exports to total exports</td>
<td>315</td>
<td>26.07</td>
<td>25.65</td>
</tr>
<tr>
<td>Log(1 + CPI inflation)</td>
<td>IFS</td>
<td>Percent</td>
<td>630</td>
<td>12.87</td>
<td>23.10</td>
</tr>
<tr>
<td>Log(1 + CPI inflation) restricted to moderate</td>
<td>IFS</td>
<td>Percent</td>
<td>574</td>
<td>6.32</td>
<td>7.74</td>
</tr>
<tr>
<td>Depreciation</td>
<td>Reinhart and Rogoff, 2004</td>
<td>Log(1 + depreciation in the parallel exchange rate)</td>
<td>459</td>
<td>8.06</td>
<td>21.25</td>
</tr>
<tr>
<td>“Open Forest” measure of export sophistication</td>
<td>Haussmann et al., 2006</td>
<td>Distance-adjusted level of income associated with all potential new export goods, where the distance is measured between each new potential good and the economy’s current export basket</td>
<td>554</td>
<td>13.35</td>
<td>0.98</td>
</tr>
<tr>
<td>Trade liberalization</td>
<td>Waciari and Welch, 2003</td>
<td>Indicator = 1 if current year is greater than the year of trade liberalization and no reversal of the trade policy reforms have occurred, and 0 otherwise.</td>
<td>484</td>
<td>0.66</td>
<td>0.47</td>
</tr>
<tr>
<td>Overvaluation of income by price level</td>
<td>PWT 6.2</td>
<td>log (PPP over GDP divided by the exchange rate times 100)</td>
<td>641</td>
<td>−18.74</td>
<td>38.10</td>
</tr>
</tbody>
</table>


Appendix B. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.jdeveco.2011.08.002.

References


Global Inequality

From Class to Location, from Proletarians to Migrants

Branko Milanovic

The World Bank
Development Research Group
Poverty and Inequality Team
September 2011
Abstract

Inequality between world citizens in mid-19th century was such that at least a half of it could be explained by income differences between workers and capital-owners in individual countries. Real income of workers in most countries was similar and low. This was the basis on which Marxism built its universal appeal. More than 150 years later, in the early 21st century, the situation has changed fundamentally: more than 80 percent of global income differences is due to large gaps in mean incomes between countries, and unskilled workers’ wages in rich and poor countries often differ by a factor of 10 to 1. This is the basis on which a new global political issue of migration has emerged because income differences between countries make individual gains from migration large. The key coming issue will be how to deal with this challenge while acknowledging that migration is probably the most powerful tool for reducing global poverty and inequality.

This paper is a product of the Poverty and Inequality Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The author may be contacted at bmilanovic@worldbank.org.
Global inequality: From class to location, from proletarians to migrants

Branko Milanovic

JEL Classification: D31, I3, N3
Key words: global inequality, inequality between nations, migration, labor in 19th century and today

1 Research Department, World Bank. The paper was written while I was a Visiting Fellow at Universidad Carlos III in Madrid.
1. Global inequality in the mid-19th century

In the spring of 1848, as the pan-European revolution was nearing its crescendo, Karl Marx and Friedrich Engels wrote probably the best known political pamphlet of all time: The Communist Manifesto. When they surveyed the situation of the world, they persuasively and repeatedly insisted on the fact that people in all “civilized” societies were divided into two large classes: that of the owners of the means of production (the capitalists), and those that were selling their labor for a living and held no property (the proletarians). It was an almost self-evident division at the level of each and every country. Having capital meant being rich; having only labor power meant being poor. There were not too many people in-between, with middling levels of income, whether those who owned some capital and yet had to work with their hands like artisans, or peasants who toiled their own land. Moreover, even they—the logic of capitalist development was implacable—were doomed to extinction or irrelevance, as they would gradually “dissolve” mostly into proletariat with perhaps only a few making it to the capitalist class. The division into two or three main classes (the third being landowners, who in Marx and Engels’ view could be assimilated to capitalists) was not, of course, introduced by Marx and Engels. It has been present in contemporary political economy, and it hailed back to Adam Smith, and even to François Quesnay. It was used by David Ricardo in his Principles, published in 1817, as a key feature, motivating his entire economic analysis.

So evident seemed the class division in all societies that Marx and Engels concluded their Manifesto by writing that “…modern industrial labor, modern subjection to capital, the same in England as in France, in America and in Germany, has stripped [the proletarian] of every trace of national character” (1978, p. 482). Being proletarian was thus a global condition, they held, and being global, it presented an ideal basis on which international solidarity of the working class could be built. Proletarians were equally poor and exploited everywhere and they could liberate themselves and usher in classless societies only in a common effort that knew no national borders. This was expressed in Marx and Engels’ famous statements “[t]he proletarians have nothing to lose but their chains. They have a world to win” (ibid, p. 500). Local emancipation and international solidarity were thus intertwined, part of the same struggle.

This same view of workers who have no true homeland because they are everywhere equally destitute and powerless was reflected, only a year earlier, in a speech delivered to the
Free Trade Congress in Brussels by Engels. He defended his and Marx’s pro-free trade stance as follows:

Under freedom of trade the whole severity of the laws of political economy will be applied to the working classes. Is this to say that we are against Free Trade? No, we are for Free Trade, because by Free Trade all economical laws, with their most astounding contradictions will act upon a larger scale, upon a greater extent of territory, upon the territory of the whole earth; and because from the uniting of all these contradictions…will result the struggle which will itself eventuate in the emancipation of the proletariat.  

Engels believed that by letting capitalist rules of the game embrace ever greater portions of the globe, the outcome of such a capitalist domination would be similarity of economic conditions among workers. The similarity in economic conditions would, in turn, lead to the concordance of economic interests and to the emergence of solidarity among workers of different countries. Ultimately, it would culminate in a worldwide revolution.

But did similarity of economic conditions among workers exist then? Did Marx and Engels depict the reality of that era correctly? Today we have more data than Marx and Engels possessed at the time. Yet, their insight is confirmed by what we know today, both as regards the ubiquitous split between the two major classes and similarity in the economic position of the laborers, or, more broadly, the poor across countries.

Consider class and income compositions of a couple of societies for which we have the data from the mid-19th century. In R.D. Baxter’s social table for England and Wales for 1867, incomes of the wage-earning classes, representing 72 percent of the population and divided into eight subgroups, start at the bottom of the distribution, just above paupers, with an estimated per capita income equal to one-third of the national mean, and end at the richest point, with an income 10 percent below the mean. The propertied classes (a little over 6 percent of the population) covered the upper part of the income distribution. The income ratio between the two classes was greater than six to one.


3 Available at Peter Lindert’s Website “Global Price and income History Group” , see http://gpih.ucdavis.edu/.
England was the prototypical and the most advanced capitalist economy. But we have the data for Chile in 1861 which reveal the same, or even sharper, polarization. Working classes, from the poorest (female fisherman) to the richest (shoe-makers), covered the range of income that goes from less than one-seventh of the national mean to two-thirds of the mean. Only the artisans, who mixed ownership of some capital with their own labor, and possibly even hired a few workers, achieved an income higher than the national average. At the high end of the income distribution were owners of manufactures and mines with incomes between 20 and 55 times the mean, and land owners and large-scale merchants with respectively 35 and 9 times the mean. These top groups comprised only 2 percent of the population while working classes (including peasants) accounted for more than 90 percent.

But if workers filled in most (or perhaps in all) countries, that part of the income pyramid which ran from the subsistence to just somewhat under the mean, their real incomes in various countries could still have differed a lot if country means were very different. But, as we shall see, around the 1850s, they were not. Angus Maddison has estimated that around 1850, the mean income in the poorest countries in the world (Ceylon and China) was around $PPP 600. At the top were the Netherlands and the United Kingdom with a GDP per capita of about $PPP 2,300. Thus, the ratio between the top and the bottom (of country mean incomes) was less than 4 to 1. Consequently, the better-off workers who earned incomes close to the national means, could not, in terms of their standard of living, differ from each other by more than the ratio of 4 to 1. And the bulk of workers who lived at less than their countries’ average income and closer to the subsistence, could not have incomes that differed by more than 2 to 1—with many of them living at approximately the same subsistence level. Indeed, Broadberry and Gupta (2006, Table 6, p. 17) show that in the period 1800-1849, the wheat-wage of an unskilled daily laborer in India (among the poorest countries in the world then) was about 30% of the wage of a similar worker in England. And comparing the Netherlands with the Yangtze valley, two relatively developed areas sharing a number of similar geographic features, Li and van Zanden (2010, p. 21) conclude

4 The data are derived from an Occupational Census, were consolidated by Javier Rodriguez Weber, and can be found on Peter Lindert’s Website “Global Price and Income History group” (see the previous footnote).

5 In 1990 international prices.
that in the 1820s, real wages in the Netherlands were about 70% higher than in the Yangtze valley.\textsuperscript{6}

Thus, similarity in the economic position of workers across the world, implicitly assumed by Marx and Engels, had a firm basis in the reality of the time. And, as we have seen, this is crucial because it is the similarity in the economic positions which allowed Marx and Engels, as well as later Marxists, to derive the principle of international solidarity of the working class.

We can reach the same conclusion that the main income cleavage was the one between classes, and not between countries, from a slightly different vantage point, if we consider global inequality, that is inequality between world citizens. In a pioneering work, François Bourguignon and Christian Morrisson (2002) have used a collection of their own and other authors’ estimates of income distributions for 33 country-groups and mean incomes from Angus Maddison to construct worldwide income distributions, at approximately twenty-year intervals, for the period 1820-1992. They estimate global inequality, measured by the Gini coefficient, to have been about 53 Gini points in 1850, and to have been composed in almost equal proportions of between-country and within-country inequalities. The former (between-country inequality) refers to that part of inequality that is due to the differences in country mean incomes while the latter (within-country inequality) is that part of total inequality which is due to inequalities existing in each individual country. To fix the ideas, we can call the between-country inequality “locational” because it depends on the differences of mean incomes between various places (countries), and the within-inequality, “class” inequality because it depends on different individuals, living in the same country, having different incomes and belonging to different social groups.\textsuperscript{7} Specifically, Bourguignon and Morrisson estimate that the global Gini in 1850, amounting to 53.2 points, can be broken down into 25.9 Gini points (49 percent) due to location, and 27.3 Gini points (51 percent) due to class. Thus, around mid-19\textsuperscript{th} century, one-half of inequality between individuals in the world was explained by unequal development of countries and another half by income

\textsuperscript{6} Li and van Zanden’s results favor the view that appreciable difference in wages existed as early as 1800s in opposition to other writers (especially Pomerantz (2001) and to some extent Bairoch, 1997, p. 111) who believe that the gaps were less.

\textsuperscript{7} Between-inequality is inequality that would exist if everybody in every country had the mean income of his/her country, or in other words, if inequality within each nation were zero.
differences between social classes—that is, essentially between workers and capitalists. How does it compare with the situation today?
2. Global inequality in the early 21st century

If we use the same decomposition between location and class today, when our data are much better than for the past, we find that of the global Gini, which amounts to 65.4 points, 56.2 Gini points or 85 percent is due to differences in mean country incomes, and only 9.2 Gini points (15 percent) to “class”. Not only is the overall inequality between world citizens greater in the early 21st century than it was more than a century and a half ago, but its composition has entirely changed; from being an inequality determined in equal measures by class and location, it has become preponderantly an inequality determined by location only. This fact is of great political and economic significance.

Figure 1 helps us visualize this new reality. On its horizontal axis we draw the population of a given country divided into twenty equally-sized groups, called ventiles, each including 5% of the population, ranked by their average per capita incomes. Thus, value 1 on the horizontal axis corresponds to the poorest 5% of the population in a given country, and value 20, to the richest 5% of the population. On the vertical axis we show the global percentile position of each national ventile. The vertical height, corresponding for example to the bottom ventile in the US, is y=60 and it indicates that the poorest 5% of Americans have an income that places them at the 60th global percentile. In other words, they are better-off than 60 percent of world population. The same interpretation applies to any other national ventile. Figure 1, using the example of BRIC countries (Brazil, Russia, India and China) and the US, illustrates vast differences in incomes which exist between countries, and in particular between the poorest ventiles of the population. While the poorest Americans are (as we have just seen) at the 60th global percentile, the poorest Brazilians and Indians are at the 3rd or 4th global percentile, that is among the poorest people on the planet. The poorest Chinese are around the 16th global percentile, the poorest Russians around the 37th percentile. Even more striking is the comparison of the income of the

---

8 All 2005 global data derived from nationally-representative household surveys conducted around year 2005 (see World Income Distribution (WYD) database available on [http://econ.worldbank.org/projects/inequality](http://econ.worldbank.org/projects/inequality)). In order to keep comparability with the precedent calculation for mid-19th century, I use price levels and PPPs that are commensurate with 1990 PPPs rather than the most recent version of PPPs (year 2005). The latter gives an overall higher global inequality (70 Gini points rather than 65.4; see Milanovic 2011, and Milanovic 2012). The share of the between-country component is also about 85 percent if we use more up-to-date PPPs.

9 The position of the poorest ventile of the rich country is indicated by a broken line in this and following graphs.
poorest groups of Americans with the richest Indians: the second poorest ventile of American population is approximately at the same level of income (just 1 global percentile lower) than the richest 5% of Indians.  

Figure 1. Importance of location vs. class: United States compared to BRIC countries, 2005

Note: Based on national household surveys; people ranked by per capita income or per capita consumption adjusted for price differences between the countries using the most recent PPPs. BRIC is an acronym denoting the emerging market economies of Brazil, Russia, India and China.

The figure could be made both more dramatic or less dramatic without losing anything of its essential message: namely, that most of global income differences today depend on location. We can make the figure more dramatic by contrasting incomes of people living in a

\[10\] Note that if class alone determined one’s global income position, and all country means were the same, the line for every country would be a straight 45 degree line.
very rich and egalitarian country like Denmark with people living in various poor African
countries (Figure 2a). Danish poorest ventile has an income that places it as the 82\textsuperscript{nd} global
percentile while in many African countries even the income of the richest ventile hardly exceeds
the 60\textsuperscript{th} or 70\textsuperscript{th} global percentile—implying thus that the two distributions do not overlap at all.
In other words, if we lined up all individuals from these countries by their per capita income,
Denmark’s income distribution would only start at the point at which many African countries’
distributions end. The richest Malians are poorer than the poorest Danes.

The picture of global location-induced inequality could be also rendered a bit less
dramatic, if we present it in a more finely-grained form, \textit{viz.}, if instead of using ventiles on the
horizontal axis, we used percentiles of national income distributions, or even individuals lined up
and ranked one by one. For sure, we should then be able to find some Malians who are richer
than some Danes, and the two distributions would indeed display some overlap. However, that
overlap would be, in a statistical sense, minimal: there may be one-half of a percept or 1 percent
or even 2 or 3 percent of Malians who are richer than the poorest Danes, but this does not in any
way invalidate the main message from our graph. In effect, when we contrast US and India, and
instead of ventiles divide their populations into percentiles, the overlap is only 4 percent: that
many Indians are better off than the poorest Americans (see Figure 2b).
Figure 2. Income gaps between various countries’ populations; early 21st century

a. Denmark vs. selected African countries (ventiles)

b. US vs. India (percentiles)

Note: Based on national household surveys; adjusted for price differences between the countries using most recent PPPs. Year 2005.
Figure 3 displays broadly the same information in a different way. It shows the distribution of per capita PPP incomes received by the lowest and highest ventiles in all countries in 2005. Two points stand out. First, the distributions are quite wide indicating that people who are nationally poor (or nationally rich) receive vastly different incomes, depending on what country they live in. Second, the two distributions do intersect, implying that there are countries such that people who are poorest there are yet better-off than the richest people in other countries. The link between being nationally and being globally poor is severed.

Figure 3. Income of people belonging to the poorest and richest country ventiles (year 2005, 114 countries)

Note: Based on national household surveys; adjusted for price differences between the countries using most recent PPPs.

The income gap between countries’ average incomes is much greater today than around 1850. Again, using Maddison’s data, in order to keep comparability with the 1850 results, we find that the top-to-bottom ratio in 2007 was in excess of 100 to 1 (as opposed to 4 to 1 that it was in 1850). The increase in the gap is easy to understand: while the poorest countries today are
not richer than the poorest countries in the past, the richest countries are immensely better-off. Instead of Britain and the Netherlands that were the richest countries in 1850, today, we have USA, Singapore and Norway, all with incomes around $PPP 30,000, that is, 13 times richer than Britain or the Netherlands in 1850. At the bottom of per capita income “league”, we have Congo, Burundi, Niger and Central African Republic, all with incomes just above subsistence, some $PPP 350-700 per capita, not at all different from the level of the poorest countries in 1850. The world today presents a peculiar picture where some of its parts are immensely richer than ever in history while other parts have an income level about the same as it was 150 or even 500 years ago. Thus to speak of an average global income is correct in an arithmetic sense but otherwise lacks much meaning.

Even when we contrast the fast-growing “emerging economies” of China and India with the rich world, the gap in the first decade of the 21st century is greater than it was around 1850. To keep to the comparison from the 1850s, consider the GDP per capita of the United Kingdom (not any longer the richest country in the world) against GDPs per capita of India and China (among the fastest growing economies in the last two decades)—all GDPs adjusted for the differential price levels between the countries. The gap between UK and India in 2009 is in excess of 10 to 1 while it was only 5 to 1 in the mid-19th century; the gap between UK and China is 5 to 1 today while in 1850 it was (as we have seen above) less than 4 to 1.

Indeed, the number of countries today is much greater than it was some 160 years ago, and it may seem that this factor alone would lead us to conclude that inter-country income differences, and the gap between the top and the bottom, must have widened anyway. But the widening is not an artifact of the number of countries. We can see this if we limit our observations to the same 63 countries as included by Maddison in 1870, and whose populations

---

11 In 1990 international prices. We leave out some outliers, small oil-producing economies and Luxemburg, with even higher incomes.

12 Maddison does not give GDP per capita estimate for India in 1850, so we use the next proximate year (1870). It is believed that Indian GDP per capita was stagnant throughout much of the 19th century, so the exact year should make little difference.

13 Data from World Bank World Development Indicators, year 2009, expressed in 2005 PPPs.

14 The number of countries included by Maddison in 1850 (twenty-four) is too small for a valid comparison.
today accounts for some 5 billion people. In 1870, the gap between the richest countries (Australia and Great Britain) and the poorest (Nepal and Ghana) was 8 to 1; in 2007, it is 31 to 1 (United States and Norway vs. Nepal, North Korea and Ghana).

Moreover, the very logic of widening income disparities alluded to before shows that the effect of the number of countries cannot be decisive. Because the logic is that some people, living in specific geographical locations, earn today approximately the same—very low—level of income as their ancestors did more than a century and a half ago, while other people, at other locations, live at income levels that are manifold greater than in the past and some 100 times above the subsistence. Thus, the ratio between top and bottom must rise, almost regardless of how we decide to “slice” the population of the world (that is, in how many countries they live).

We obtain the same conclusion if we compare real wages in rich and poor countries. Table 1 shows wages for three occupations in five selected countries (or rather their major cities): two rich (London and New York) and three poor (Beijing, Delhi and Nairobi). The data refer to after-tax nominal wages which are then deflated by the food cost of living in the same cities to obtain real food wages. The advantage of these data, collected for 74 cities and 14 occupations and published by Union de Banques Suisses, UBS (2009), is their almost full comparability. In effect, the wage data refer to net wages earned by the same, very narrowly defined, professions (see Notes to Table 1); they are adjusted for the effective number of hours worked, and are geographically limited to large cities for which UBS publication also calculates a food cost of living index.

In Table 1, I have selected three occupations with increasing level of skills: building laborer, skilled industrial worker, and engineer. The real wage gaps are the greatest in the case of unskilled workers: for them, the ratio of real wages in rich and poor countries is almost 11 to 1.

---

15 The caveat “almost” is necessary because in the extreme case, when the world population were divided into (say) two countries and in such a way that each half would contain the same proportion of locations that have not seen any progress since 1850 and locations that are today much richer, the gap between these two halves need not be greater, and may be smaller, than it was in 1850. But obviously, this is an extreme case which has nothing in common with the real processes of state-formation and real state of affairs both now and in the past.
For the skilled industrial workers and engineers, the gaps are respectively 5.8 to 1 and 3.3 to 1.\textsuperscript{16} If we contrast this result with the estimates of the ratio between wheat wages of English and Indian unskilled laborers around 1850 (Broadberry and Gupta, 2006, quoted above), we note that the UK/India unskilled wage gap has increased from around 3.3 to 1 in 1850 to more than 9 to 1 today. The advantage of this particular comparison is that it keeps the level of skills constant across time, and focuses on the very low skills representative of those who are at the bottom of the income pyramid.

It is the gap between the poorest members of rich and poor societies that we found to be the widest, both in income and wage data. Or, to translate this finding in terms of political economy, the commonality of interests between the poor and unskilled workers in the rich world and the poor and unskilled workers in the poor world is hard to detect if we focus on their economic conditions only. This is of course a major departure from a situation which existed a century and a half ago.

\textsuperscript{16} A decreasing rich-poor country wage gap with greater skills of workers is also found by Warner (2002). Warner attributes it to greater global competition which exists for more skilled workers whose wages reflect international conditions while the wages of low-skilled workers behave like non-tradables. It can also be explained by greater relative scarcity of more skilled workers in poorer countries.
Table 1. Nominal and real (food) hourly wages for several occupations
Annual after-tax wage divided by the number of effective annual hours of work, March 2009

<table>
<thead>
<tr>
<th></th>
<th>Building laborer 1/</th>
<th>Skilled industrial worker 2/</th>
<th>Engineer 3/</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nominal after tax</td>
<td>Real food wage</td>
<td>Nominal after tax</td>
</tr>
<tr>
<td>wage (in $)</td>
<td>wage (in $)</td>
<td>(in $)</td>
<td>wage (in $)</td>
</tr>
<tr>
<td>New York</td>
<td>16.6</td>
<td>16.6</td>
<td>29.0</td>
</tr>
<tr>
<td>London</td>
<td>9.7</td>
<td>15.4</td>
<td>19.0</td>
</tr>
<tr>
<td>Beijing</td>
<td>0.8</td>
<td>1.3</td>
<td>2.3</td>
</tr>
<tr>
<td>Delhi</td>
<td>0.5</td>
<td>1.7</td>
<td>2.1</td>
</tr>
<tr>
<td>Nairobi</td>
<td>0.6</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>Rich vs. poor (unweighted ratio) 4/</td>
<td>20.4</td>
<td>10.9</td>
<td>11.0</td>
</tr>
</tbody>
</table>

Note: Food prices are estimated from a basket of 39 food products with weights reflecting West European consumption patterns. New York food prices are set equal to 1. Real food wage (in New York food prices) is estimated by dividing the nominal after-tax dollar wage by the food price index (not shown here). Annual number of hours worked is equal to the weekly number of hours of work given for each profession and country separately (UBS, 2009, pp. 34-35) multiplied by 52 weeks, and reduced for the number of official and paid vacation days per year for each country (UBS, 2009, p. 30).

1/ Unskilled or semi-skilled laborer, about 25 years of age, single.
2/ Skilled worker with vocational training and about 10 years of experience, working in a large company in the metal working industry, approximately 35 years of age, married, two children.
3/ Employed in an industrial firm in the electrical engineering sector, university or technical college graduate with at least 5 years of experience, about 35 years of age, married, two children.
4/ Rich are New York and London; poor are Beijing, Delhi and Nairobi.
3. From “permanent revolution” to “fortresses Europe and America”

The fact that the position of the poorest members of different societies is vastly different—the fact, for example, that the poorest 5% of Americans, adjusted for price levels, earn 35 times more than the poorest Zambians or 12 times more than the poorest Malians—has global political implications. We have seen above that the entire construct of workers’ (poorest people) solidarity was based on some objective conditions, that is on similarity in their living conditions. But if that similarity in living and working conditions has evaporated, would not the commonality of interests and “class solidarity” similarly vanish? And indeed it is not easy to find any examples of shared interests between the poorest classes in rich countries and poorest classes in poor countries. More likely, their interests conflict.

A second implication of the world where location determines to a large extent one’s income is that it must be a world of huge migratory pressures because people can increase their incomes severalfold if they migrate from a low mean-income location to a high mean-income location. Only if knowledge of these income differences could be somehow hidden, and not widely-shared, could we expect that people would not want to migrate. But this is most patently not so in the era of globalization, instant communication, and broad access to TV, movies and Internet.

These two factors thus shape the key global political issues. While in the years between the heady days of the pan-European “Springtime of peoples” in 1848, and probably the second-half of the 20th century, the conflict between capital and labor was the main political issue that influenced several generations of thinkers, politicians, social activists, and ordinary people, this is no longer the case today. Globally, the issue has receded in importance as the objective conditions that gave rise to it have changed. This was already even if dimly becoming apparent in the last decades of the 19th century when the term “workers’ aristocracy”, denoting this divergence of living conditions among the “exploited” classes internationally, was coined. To quote Engels again—but now ten years after *The Communist Manifesto*: “…the English proletariat is actually becoming more and more bourgeois.”

17 In a letter to Marx in 1858, *Marx and Engels: Selected Correspondence*, p. 132. See also Hobsbawm (1996, pp. 224-229).
This somewhat derogatory term reflected a real process of betterment of the standard of living among the working classes in the most advanced capitalist countries, or if one prefers, their “embourgeoisement”. The process continued, and accelerated, most famously during the “Glorious Thirty” years of almost uninterrupted and widely shared economic growth in Western Europe and the United States. In the second half of the 20th century, the prospect of permanent revolution, that is of a worldwide revolution that would bring working classes’ parties to power—a prospect so eagerly called forth by Trotsky in the 1920s, became clearly unrealistic. The bourgeoisies of the western world needed no longer fear proletarian revolutions spreading from one country to another. The fact that in 1968, during another pan-European uprising, superficially similar to that 120 years earlier, workers’ organizations in the most “revolutionary” country (France) were the last to join the protest, unable to formulate any coherent demands, and in fact never made any moves towards the abolition of capitalist relations of production, showed very clearly that the times had changed. The specter of Communism, eloquently evoked by Marx and Engels in the opening sentences of the Communist Manifesto, was exorcised.

The new problem which is likely to dominate the present century is different: it is the problem of uneven development between the countries and, associated with it, the pressure of migration emanating from poor countries. It is no less complicated problem than the earlier one, and involves similar fears of loss of power and income among those who are richer. It is a problem born out of importance of location for one’s lifetime earnings no less than the earlier fear of Communist revolution was born out of income differences between the social classes within the same societies.
4. The key policy issue in this century: Better living standards for the poor people of the world

The specter of Communism disappeared because poor people in what are today rich and upper middle income countries have experienced substantial and sustained increases in their real levels of incomes. Analogously, the problem of migration will disappear, or become manageable—in the sense that migration will be just a result of people’s preferences (e.g. those who prefer to live in a warmer country may relocate there, as is broadly the case today among the group of rich European countries) and not a problem of massive exodus, prompted by large gains in real income—only when differences in mean incomes between countries become much smaller. This implies that the best “solution” to the problem of massive, economically-motivated migration lies in the accelerated development of the poorest countries of the world and that it is also in the well-understood (self-enlightened) interest of rich nations to help this process by increased aid and economic benefits given to the poor countries. The self-enlightened interest should be particularly obvious in the case of countries that have problems managing large migration flows whether because of economic reasons (possible downward pressure on wages) or social reasons (difficulty of accepting a different style of life or a different system of values brought in by the migrants).

But the development of many poor parts of the globe, even if it were more successful than during the last 60 years, would still take a long time to bring their average incomes to anything close to the incomes in the rich world. Thus migration will remain, by its sheer human importance, the key mechanism whereby incomes of the poor people in the world are to be raised. Faster development of poor countries and migration are two complementary, and in many ways very similar, mechanisms. In both cases, the end result is increase in income of the poor. In one case, the increase is achieved in the place where they live now; in the other case, it necessitates their movement to a different locale. The process moreover is hardly new: in terms of relative importance, migration was much more significant at the end of the 19th century.

18 The same point is argued by Pritchett (2006, Chapter III in particular).
and early 20th century than it is today. Not because the demand for migration is less today but because the impediments are greater.

Impediments to migration pose one methodological issue for economic theory, a point raised by Edwin Cannan almost a century ago, in his discussion of Adam Smith’s “invisible hand”. Cannan asked “if…indeed, it [is] true that there is a natural coincidence between self-interest and the general good, why…does not this coincidence extend, as economic processes do, across national borders?” Smith’s argument of “invisible hand” is general and cannot depend on arbitrary lines such as national borders. The relocation of people ought to be beneficial to total world output and therefore to reduction of global poverty and (very likely) global inequality. Hanson (2010, p. 22) calculates that the annual flow of Mexican migrants into the United States raises global income by an amount equal to about 1% of American GDP. Walmsley and Walters (2006) find, using CGE simulations, that an increase in both skilled and unskilled migration equal to 3 percent of their respective labor forces in developed countries, would yield severalfold more, in net welfare gains terms, than the current development aid. Migration restrictions are, they argue, more costly than the existing restrictions on trade. In a recent study of the effects of migration in Spain, Bruquetas Callejo and Moreno Fuentes (2011) find that the immigration surge that has, in less than two decades, transformed the country from an emigrant nation to one where foreigners account for 12% of the resident population, has benefited Spain.

Pritchett (2006, p.95) criticizes economists who, while acknowledging that both trade and migration are welfare-enhancing, tend to treat them very differently. They argue for trade liberalization on general welfare grounds, and then, in a second step, address income distribution concerns of those who are affected through redistribution. But they never use the same approach

---

19 Pritchett (2006, p.69) estimates that the net flows (relative to population) are today between ½ and 1/5 of what they were in the period 1870-1910.

20 Cannan’s question is quoted from Frenkel (1942, p. 177), I am grateful to Tony Atkinson for bringing to my attention this undeservedly obscure reference.

when it comes to migration: the equivalence would imply being in favor of migration as a default position, and mitigating its negative effects through specific additional policies.

That from the global perspective migration should be desirable leaves very little doubt, even when we think of it using the simplest economic principles: if people are allowed to move where they think they would do better, it is very likely that total output would increase compared to the situation when people are not allowed to move. For if the reverse were the case, impediments to migration similar to the ones that currently exist at the international level would be found useful and imposed at the national level as well. As Frenkel (1942, p. 183) writes, “The movement of men and women from areas where they are poverty-stricken to areas where they can make their full contribution to the world’s income stream is of advantage to all”.

While the desirability of both increased aid and greater migration may be established in principle and for the world as a whole, it does not mean that it would be to the advantage of each particular country, or particular sections of population in each country. Greater migration may be associated with reduced wages or increased unemployment for the groups of people whose skills are most similar to the skills of migrants. Thus even if for the recipient country as a whole migration is advantageous, sufficiently powerful economic and political groups may be able to block it or impose tough limits on it. Yet, as nicely put by Hanson (2010, p. 22), “in considering the migration rights that maximize global welfare, one could not argue that US workers be weighted more heavily than both richer [US] employers and poorer migrants”. Again, going back to the first principles, we can either treat every person in the world as equally important, or we can put a greater weight on the welfare of poorer individuals because of diminishing marginal utility of income but we cannot (logically) pick and choose one group as such in preference to others.

Perhaps more importantly, greater migration faces non-economic obstacles that can be vaguely described as belonging to the area of social acceptance of migrants and their effects on domestic culture. These issues have been exacerbated by the current economic crisis and the unexpectedly great difficulties that many European countries have had in “absorbing” migrants, particularly those with Islamic background. Thus, in a close succession, both British and German Prime Ministers have declared the “multiculturalist” model, which was supposed to be Europe’s answer to migration, to have failed. Angela Merkel, moreover, pronounced such a model unambiguously “dead”. These problems hold particular poignancy for Europe because it is, due
to its low fertility rates, much more in need of migrants than the United States. Moreover, in its
southern fringes, Europe is surrounded by countries whose demographic profiles are exactly the
opposite of the European: these are countries with a very large share of young population and
relatively few older people. Were it not mostly for non-economic reasons, the demographic
match between the two shores of the Mediterranean would seem almost perfect. However,
Europe which has, at least since the end of the 15th century, “exported” its people elsewhere has
a serious difficulty viewing itself as an immigrant continent and accepting migrants who mostly
come from the areas that were formerly colonized by the Europeans. This reversal of migration
flows seems politically difficult to manage.

From the facts that (i) most of today’s global income inequality is due to differences in
mean incomes between the countries, (ii) in an era of globalization such differences are well-
known to people in poor countries, and (iii) the costs of moving from one place to another are
not prohibitive, it follows that migration, in the absence of significant acceleration of growth in
poor countries, will be a great 21st century mechanism of “adjustment”. It will be driven by the
self-interest of individuals but its ultimate result would be a reduction in global inequality and
global poverty. Aid and migration ought to be regarded as two complementary means for
achieving these goals. Policy makers in developed countries shall come to realize that either
poor people will become richer in their own countries or they will migrate to the rich countries.
REFERENCES


This paper was prepared for the Handbook of the Economics of European Integration edited by Harald Badinger and Volker Nitsch, to be published by Routledge. I thank the editors and anonymous reviewers for their comments. All errors are mine. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

NER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2015 by Enrico Spolaore. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.
The Political Economy of European Integration
Enrico Spolaore
NBER Working Paper No. 21250
June 2015
JEL No. F02,F15,F5,H41,H56,H77

ABSTRACT

This paper discusses the process of European institutional integration from a political-economy perspective, linking the long-standing political debate on the nature of the European project to the recent economic literature on political integration and disintegration. First, we introduce the fundamental trade-off between economies of scale associated with larger political unions and the costs from sharing public goods and policies among more heterogeneous populations, and examine the implications of the trade-off for European integration. Second, we describe the two main political theories of European integration - intergovernmentalism and functionalism - and argue that both theories capture important aspects of European integration, but that neither view provides a complete and realistic interpretation of the process. Finally, we critically discuss the actual process of European institutional integration and its limits, from its beginnings after World War II to the current crisis.

Enrico Spolaore
Department of Economics
Tufts University
Braker Hall
8 Upper Campus Road
Medford, MA 02155
and NBER
enrico.spolaore@tufts.edu
1. Introduction and Summary

The process of European integration is based on a time-honored strategy of partially integrating policy functions and institutions in a few areas – such as coal and steel, trade, or, later, a common currency - with the expectation that more integration will follow in other areas over time.

This strategy became the main approach to European institutional integration in the 1950s, after the collapse of a more ambitious attempt to create a defense and political community, which would have included a common army, a common budget, and common legislative and executive institutions - basically, a European federation. Faced with the failure to form a full political union directly, supporters of European integration pursued an alternative path of gradual and partial integration. The process took place mostly in technical and economic areas but with the expectation that deeper, more “political” integration would follow, in part as a result of the pressure from inefficiencies and crises associated with incomplete integration. From this perspective, incompleteness was not seen as a bug but, possibly, as a feature, as it was expected to lead to further integration down the road.

This gradualist strategy was effective at fostering durable integration when applied to areas with large economies of scale and relatively low costs from heterogeneity of preferences and traits across different populations – for example, the creation of a common market. The approach, however, also led to the creation of incomplete and inefficient institutional settings. Most notably, the euro was introduced in the absence of other institutions historically associated with a successful monetary union, resulting in a “half-built house” (Bergsten, 2012). A widespread rationalization of the imperfections and shortcomings of European institutions was based on the expectation that the problems associated with previous steps could always be fixed by more integration: commercial integration and monetary integration would in due course be followed by more institutional and political integration, such as a banking union, a fiscal union, or even a fully-fledged political union, in what has been described as a “chain reaction” towards an “ever-closer union.”

A fundamental issue with this chain-reaction approach is that it tends to underestimate the costs and constraints associated with heterogeneity of traits and preferences over public goods and policies in populations with diverse societal structures, cultures and identities. In fact, the trade-off between benefits from integration and heterogeneity costs is at the center of a vast and growing literature on the political economy of integration (and disintegration). An analysis of the implications of such trade-off can shed insights on the actual process of
European integration and its limitations, as well as on the theories that have been developed over the decades to understand the objectives and strategies behind the European project.

This paper provides a discussion of the process of European institutional integration from a political-economy perspective, linking the long-standing political debate on the nature of European integration to the more recent economic literature on political borders.¹

The rest of this paper is organized in three parts. Section 2 briefly discusses a few key concepts on the political economy of integration and their implications for the European project, with an emphasis on the basic trade-off between economies of scale and scope associated with larger political unions and the costs from sharing public goods and policies among more heterogeneous populations. Section 3 covers the two main political theories of European integration, intergovernmentalism and functionalism, and argues that both theories capture important aspects of European integration, but neither approach provides a complete and realistic interpretation of this complex process. Finally, the actual process of European institutional integration, from its beginnings after World War II to the current crisis, is discussed in Section 4. Section 5 concludes.

2. The Political Economy of Institutional Integration: A Fundamental Trade-off and Its Implications for Europe

A useful starting point to study the political economy of institutional integration is the fundamental trade-off between economies of scale and heterogeneity costs.² When larger and diverse groups form common institutions and pool public functions and policies - a common legal and judicial framework, a common currency, fiscal policies, defense and security, and so on - they can benefit from economies of scale in the provision of public goods, which are non-rival in consumption and therefore cheaper on a per-capita basis when the costs are spread over a larger population. Larger jurisdictions may also allow governments to internalize externalities over a broader area, and to provide insurance against shocks, such as natural disasters or regional economic crises.

Larger and more diverse jurisdictions, however, tend to face higher heterogeneity costs stemming from different political, economic and cultural traits and conflicting preferences

¹ This paper paper was prepared for the Handbook of the Economics of European Integration edited by Harald Badinger and Volker Nitsch, to be published by Routledge, and heavily builds on Spolaore (2013). General discussions of the economic approach to political borders and integration are provided in Alesina and Spolaore (2003) and Spolaore (2006, 2014). For overviews of the large literature on European institutional integration by political scientists and political economists, see for example Gilpin (2001, chapter 13), Eichengreen and Frieden (2001), Eichengreen (2006 and 2012), and Sadeh and Verdun 2009. Recent historical studies of the process of European integration include Gilbert (2012), Ludlow (2006), James (2012), and Mourlon-Druol (2012).
over public goods and policies. The relations between various measures of heterogeneity (ethnic and linguistic fractionalization and polarization, measures of genetic and linguistic distance, and other measures of historical and cultural diversity) and a series of political outcomes (provision of public goods, quality of government, redistribution, conflict within and across states) have been documented in a vast and growing empirical literature, including for instance Alesina et al. (2003), Alesina and La Ferrara (2005), Montalvo and Reynal-Querol (2005), Esteban, Mayoral and Ray (2012), Desmet, Ortuño-Ortín and Wacziarg (2012), Spolaore and Wacziarg (2012), and Arbatli, Ashraf and Galor (2013). This empirical literature is still in its infancy, and there are numerous open questions about the definitions and methods to measure the effects of heterogeneity on different outcomes. Nonetheless, a general finding is that more heterogeneous populations typically face higher political costs in the provision of public goods and a higher likelihood of civil conflict.

In principle, heterogeneity can also be a source of benefits as well as of costs for societies. In communities where agents have diverse preferences and characteristics, individuals and groups can benefit by specializing in the production of different goods and services, while also learning new ideas from each other. Benefits from heterogeneity, however, are mostly about interactions over rival goods, which cannot be consumed simultaneously by several people. In fact, low heterogeneity may lead to conflict if different individuals and groups share very similar preferences over the same rival goods, such as specific territories and resources (Spolaore and Wacziarg 2012). The opposite relation holds for diverse preferences over non-rival goods - such as a common government, legal system, and public policies - which must be shared by all within a given political jurisdiction, whether they like them or not. In the area of public goods, therefore, different preferences mean higher political costs and a higher likelihood of domestic conflict. In sum, heterogeneity of traits and preferences is mostly beneficial when different individuals and groups interact about rival goods but costly when the interaction is about non-rival goods. Consequently, heterogeneity of preferences over public goods is a major limit to the integration of institutions that provide common policies to large and diverse populations.

Up to a point, the trade-off between economies of scale and heterogeneity of preferences can be addressed through decentralization at different layers of administration. Public goods with higher heterogeneity and lower economies of scale can be more efficiently provided at lower administrative levels (e.g., municipal and regional governments), while public functions with higher economies of scale and externalities, relative to heterogeneity costs, can be centralized.

---

3 In particular, an important question is the extent to which ethnic and cultural divisions are endogenous, and respond to political and institutional change - an issue on which we will return in Section 4.
at increasingly higher levels. These ideas are partly reflected in the legal documents at the basis of European integration. For instance, the Maastricht Treaty of 1992, which reorganized and redefined European institutions, specified the following “principle of subsidiarity” (Article 3b): “In areas which do not fall within its exclusive competence, the Community shall take action, in accordance with the principle of subsidiarity, only if and in so far as the objectives of the proposed action cannot be sufficiently achieved by the Member States and can therefore, by reason of the scale or effects of the proposed action, be better achieved by the Community.”

Nevertheless, the creation of layers of distinct and overlapping jurisdictions, each organized around a subset of public functions and policies with different economies of scale and heterogeneity costs, faces limitations and challenges both in terms of economic efficiency and political stability. A key issue for disentangling different functions at different levels is that the provision of public goods comes not only with significant economies of scale, but also with *economies of scope*. It is usually more efficient to provide several public goods together, rather than through separate authorities. A particularly important kind of “economies of scope” is associated with the exercise of fundamental *sovereignty* and monopoly of *legitimate coercion*, which is a prerequisite for the provision of a vast range of public goods and policies. In fact, even when different public goods are decentralized at lower administrative layers, modern federal systems tend to centralize sovereignty – which can be defined as the residual power to take fundamental decisions over domestic and foreign relations for a state or federation, after all other specific rights and powers have been assigned to various layers of authorities.

In practice, the centralization of sovereignty is often achieved through the pooling of the means on which the ultimate monopoly of legitimate coercion depends, including explicit military power. Consequently, the formation of a sovereign polity usually goes hand in hand with the integration of defense and security under one authority, which exercises the ultimate monopoly of coercion within a territory. In turn, such power of coercion can be used to collect resources and finance a broader set of public goods, on which different groups and individuals may have different preferences. Therefore, ultimate political integration - the formation of a sovereign state or federation - has historically been associated with the pooling of defense, security and foreign policy – public functions with very high economies of scale and scope but also very high heterogeneity costs across large and diverse

---

4 These issues are at the center of the large literature on fiscal federalism - e.g., Oates, 1999; for a discussion from a political-economy perspective see Alesina and Spolaore, 2003, chapters 2, 9 and 12.

populations. It is not clear how the European “principle of subsidiarity” would apply to the fundamental issue of centralizing sovereign power and control over means of coercion.

Over the centuries, the formation of large and heterogeneous states, federations and empires has taken place as the result of actions by non-democratic rulers (Leviathans) interested in maximizing their own rents while ignoring the preferences of large part of their subjects, and/or in response to significant external security threats. In contrast, there are very few historical examples – if any – of consensual formation of sovereign states or federations by large and diverse populations under democratic and peaceful conditions.

In this respect, the history of European institutional integration so far has been no exception. A founding document of the process of European integration is the Schumann declaration of 1950, which defined the pooling of coal and steel production as “the first concrete foundation of a European federation indispensable to the preservation of peace.” Nonetheless, no European federation was formed. In 1952 the six founders of the European Steel and Coal Community signed a treaty for the establishment of a European Defense Community and a European Political Community, which would have included a common army, a common budget, and common legislative and executive institutions - basically, a European federation. The project was abandoned, however, after the treaty failed to be ratified in the French parliament. Instead, the supporters of European integration focused on the creation of a European common market, established with the Treaty of Rome in 1957. The Treaty of Rome no longer mentioned a European federation, but stated the vaguer objective of laying the “foundations of an ever-closer union among the peoples of Europe.”

Consistently with the analysis of this Section, those earlier successes of European integration took place in areas, such as commercial integration, where economies of scale are very high, while heterogeneity costs are relatively low and partially offset by benefits from diversity. In contrast, failure to integrate occurred in areas, such as defense and security, with the highest political costs from heterogeneity. The subsequent history of European institutional integration stems from the lessons (both learned and not learned) of those early successes and failures at integration, as we will see in the rest of this chapter.

---

6 For a classic analysis of federalism from this perspective, see Riker (1964). Conflict and political borders are studied in Alesina and Spolaore (2005, 2006) and Spolaore (2012). For a discussion of the attempts to integrate defense and security in Europe see Spolaore (2013, pp. 128-131).
3. Political Theories of European Integration: Intergovernmentalism vs. Functionalism

Europe’s political economy is notoriously complex and controversial. Over the decades, scholars and commentators have emphasized different motivations, strategies and interactions among the several actors involved in the process of European integration, from national governments and voters to supranational technocrats and domestic interest groups.

A traditional distinction in the political literature on European integration is between the “intergovernmentalist” view and the “functionalist” view. The two views differ in their answers to two basic questions: what are the objectives of European integration, and who is in charge?

Intergovernmentalists believe that the European project is in the hands of national governments who pursue domestic interests, mostly in the economic area. For instance, Moravcsik (1993, 1998, 2012), a leading proponent of this theory, argues that national governments have built European institutions to pursue the economic interests of their domestic constituencies, and views the euro as an economically-motivated project, mainly reflecting the interests of German exporters and other powerful economic agents. This line of analysis is part of a broader political-economy literature stressing the connections between domestic economic interests and national attitudes and policies towards European integration (for example, Frieden 2002).

From an intergovernmentalist perspective, the European Union is just a particularly complex international organization of sovereign states. European supranational institutions, such as the European Commission and the European Court of Justice, are only instruments and commitment devices that nation states have built and use in order to pursue their own objectives, while their national governments retain all fundamental power about key decisions.

---

7 We only focus on these two views here because, historically, they have been the most influential in the study of European integration by political economists. Of course there exist many other political theories and interpretations of European integration - for example, social constructivism, which investigates how identities such as “European citizenship” have been “socially constructed” through the use of norms and language (see for example Rosamond, 2003) - a theme connected to the endogenous formation of a European identity, on which we briefly touch below (see also the discussion in Spolaore, 2013).

8 The intergovernmentalist view of European integration is sometime qualified as “liberal” intergovernmentalism to distinguish it from “realist” approaches that also emphasize the central role of nation states, but stress power and military interests rather than domestic economic goals achieved through international cooperation (e.g., Garrett 1993 and Gilpin 2001).
Functionalists, in contrast, believe that supranational institutions are distinct from national governments, and fully in charge of specific functions - hence the term “functionalism.” In their view, “supranational actors” such as Jean Monnet (head of the European Coal and Steel Community in the 1950s) or Jacques Delors (head of the European Commission in the 1980s and early 1990s) played an autonomous role and provided impetus to the process, independently of national governments’ more parochial interests.

Perhaps even more important is the functionalists’ different emphasis on the long-term objectives and dynamics of European integration. The process of functional integration, while starting within specific and relatively narrow economic functions (coal and steel, trade), is expected to move to broader and more “political” areas. In the long run, economic integration is expected to lead to political integration – either to a fully-fledged sovereign federation (“the United States of Europe”) or to a “post-modern” political community in which traditional sovereign states have become much less powerful or even obsolete.

From a functionalist perspective, building partial and incomplete institutions is not a shortcoming, but a natural feature of a dynamic process, whereas crises and problems associated with previous integration can be solved through further integration. The creation of the euro can then be seen as the ultimate example of functionalist integration. From this perspective, commercial integration and capital mobility could be maintained within Europe only by forming a monetary union, which, in turn, might be sustained in the long run only through further institutional integration: a banking union, a fiscal union, possibly full political unification.

Both the intergovernmentalist view and the functionalist view capture important aspects of European integration. However, neither view provides a complete, realistic and satisfying interpretation of this complex process.

Intergovernmentalists are fundamentally correct when they stress the central role of national governments and national interests in the actual process of European integration. The history of the European Union (and of its predecessor, the European Community) shows that Europe’s supranational institutions – such as the European Commission or the European

---

9 The leading functionalist theorist of European integration was Haas (1958, 1964). The view of Haas and his followers is sometime labeled as “neo-functionalist,” to distinguish it from Mittrany’s pre-existing theory of international integration (Mittrany 1975). A critical reconsideration of functionalism was provided by Haas (1975). For more recent analyses from a functionalist perspective see for instance Pierson (1996), Sandoltz and Stone Sweet (1998), and Stone Sweet (2000).

10 For an early critical discussion from a political perspective see Hoffman (1966).

11 Recent historical studies, such as Gilbert (2012), Ludlow (2006), James (2012), and Mourlon-Druol (2012), have also moved away from a stark dichotomy between functionalist and intergovernmentalist perspectives, while embracing more complex and nuanced interpretations of the actual process of European integration.
Court of Justice - cannot move far against the fundamental interests of national governments. Supranational institutions and procedures, while playing an important role in the daily functioning of Europe, “could not work for a week in the absence of the will to cooperate of the member states, especially the largest ones – Germany and France above all” (Gilbert 2012, p. 3).

Moreover, by emphasizing and analyzing the interactions between national governments and domestic economic interests, this line of research provides useful insights on the details of the political economy of European integration.\(^\text{12}\)

Intergovernmentalist analyses, however, with their sharp focus on domestic economic interests, can miss the central role of long-term political goals and strategies that have historically determined the process of European integration. At its roots, the European project is motivated by broader political considerations. Since its very beginnings, right after World War II, the paramount goal of the European project has been to build, through gradual integration, a political and institutional system that would prevent the tragedies of the first half of the 20\(^{th}\) century, when independent and unconstrained nation states had pursued unilateral and costly protectionist policies during the Great Depression and engaged in two enormously destructive wars. Even though economic interests have certainly played an important role in the actual process of European integration, the overall design and strategy would not be comprehensible without considering its long-term political and strategic motivations.

Insofar as it emphasizes the ultimate political goals of the process and its dynamic aspect, functionalism is much closer to capture the political and ideological framework and strategy behind much of the European construction, from the earlier steps by Jean Monnet and his followers in the 1950s to the new impetus provided by Jacques Delors and his collaborators in the 1980s and 1990s. Nevertheless, the functionalist interpretation is not fully adequate to capture the actual process of European integration either. A problem with the theory is that it tends to overestimate the role and powers of supranational agents and institutions - in this dimension, as already mentioned, the intergovernmentalist view is much more realistic.\(^\text{13}\)

The overestimation of supranational actors is a symptom of a deeper issue with the functionalist view, stemming in part from its dual nature, as a positive description of the ideology and goals of the supporters of European integration, but also as a normative theory of how European integration should actually proceed. While the functionalist analysis is a very useful description of the political objectives of European integration, it is also an

\(^{12}\) For example, see Frieden (1998, p. 33) for prescient insights on the political economy of European integration and the euro.

\(^{13}\) For a more extensive discussion of this point, see Spolaore (2013, pp.136-138).
involuntary mirror of its problems and limits. As a normative strategy of integration, the functionalist approach tends to underestimate the obstacles and limitations that would eventually affect a dynamic process of gradual integration towards an “ever-closer” union.

As highlighted in the previous section, in order to understand the political economy of European integration it is crucial to consider the implications of the fundamental trade-off between benefits from integration and heterogeneity costs. A central problem with the functionalist strategy of European integration, based on gradual integration of specific functions, has indeed been the lack of a realistic assessment of the increasing costs and constraints imposed by heterogeneity of preferences over the provision of public goods and policies when populations have different traits, cultures and identities.

As we already mentioned in Section 2, stable integration is more likely to take off in areas such as commercial integration, where heterogeneity costs are relatively low and partly offset by the benefits from diversity. As integration proceeds to other areas, heterogeneity costs continue to increase and become politically prohibitive. The functionalist approach does not anticipate that heterogeneity costs and constraints will eventually become binding and may stop the process for good. Followers of this approach are therefore prone to setting up incomplete and inefficient arrangements, relying on the overoptimistic expectation that such inefficiencies can always be addressed at a later stage through additional integration.

Hence, as we will see in more detail in the next section, the functionalist perspective can provide a useful interpretation of the earlier successes towards European integration, but its optimistic implications are not a reliable guidance to the problems and limitations that the process is bound to meet as functional integration moves from lower-heterogeneity areas to higher-heterogeneity areas.

4. The Actual Process of European Integration and its Limits

The history of European institutions includes policies that resulted in durable integration (“successes”) as well as policies that did not (“failures”). In fact, such history started with an early success (the formation of a coal and steel community, proposed with the Schuman declaration in 1950 and established with the Treaty of Paris in 1951), and a dramatic failure (the collapse of the defense and political community in 1952, discussed in Section 2). Those two different experiences motivated the subsequent strategy of European integration. The fathers of European institutions came to believe that the creation of an ambitious federal structure with major political functions faced insurmountable political obstacles, at least in the shorter run. Instead, they hoped to proceed towards increasing political integration through a gradualist and dynamic strategy, basically along the functionalist lines described in the previous section. Specific functions could be delegated to supranational institutions in
relatively narrow areas, mostly technical and economic (coal and steel, common market, later a common currency), but with the expectation that this would lead to more institutional integration in other areas over time. In other words, Monnet and his followers shared the functionalist view that partial integration would gradually lead to an ever-closer union over time, by creating pressure for more functional integration.

The strategy was partly motivated by the hope that, in the long run, national politicians, voters and interest groups would learn about the benefits of integration, and would therefore demand broader and deeper integration in more areas. Even more important was the expectation that different European populations and policy-makers, by learning to interact and cooperate in economic and institutional matters, would gradually converge in values, norms, and preferences. Over time, this would lead to an “endogenous” reduction in what we have called heterogeneity costs, therefore facilitating further integration in more sensitive and political areas.

These optimistic expectations have been fulfilled only in part. There is no doubt that the European project has greatly expanded from its early beginnings, dramatically growing both in member countries - from the initial six to twenty-eight - and in the extent of functions involved - from a European Coal and Steel Community (ECSC) to a European Economic Community (EEC) to a European Union (EU) including an economic and monetary union (EMU).

It is however much less clear whether European policy-makers and populations have converged in values and preferences over public goods and policies, along the lines expected by the supporters of European integration. In principle, cultural traits and preferences can indeed change and adapt in response to economic and political changes. There is little evidence, however, that the kind of economic and political cooperation associated with the building of European institutions is bringing about a significant convergence in national preferences, characteristics, and behaviors or the formation of a unitary “European identity.”

Nevertheless, the builders of European institutions did not rely only on positive mechanisms, such as an increasing demand for integration due to learning and convergence of preferences. Monnet and his followers also expected that partial integration might lead to further integration, paradoxically, because of its own shortcomings and limits – its own

---

14 For example, see Fearon (2006) for a discussion of the political literature on how ethnic and linguistic divisions and their relevance can be affected by political and institutional changes. Bisin and Verdier (2010), Spolaore and Wacziarg (2013) and Spolaore (2014) provide discussions of the growing economic literature on culture and economic outcomes.

15 Analyses of the political economy of “nation-building” are provided in Alesina and Spolaore (2003, pp. 76-78) and Alesina and Reich (2013).
“incompleteness.” This was clearly explained by one of Monnet’s collaborators (Ball, 1994, p. 10):

“There was a well-conceived method in this apparent madness. All of us working with Jean Monnet well understood how irrational it was to carve a limited economic sector out of the jurisdiction of national governments and subject that sector to the sovereign control of supranational institutions. Yet, with his usual perspicacity, Monnet recognized that the very irrationality of this scheme might provide the pressure to achieve exactly what he wanted - the triggering of a chain reaction. The awkwardness and complexity resulting from the singling out of coal and steel would drive member governments to accept the idea of pooling other production as well.”

More recently, the functionalist argument that partial steps in integration would create need and pressure for further integration, in a sort of “chain reaction,” was explicitly echoed by Tommaso Padoa-Schioppa (2004, p. 14), a close collaborator of Jacques Delors and a key architect of the euro:

“[T]he road toward the single currency looks like a chain reaction in which each step resolved a preexisting contradiction and generated a new one that in turn required a further step forward. The steps were the start of the EMS [European monetary system] (1979), the re-launching of the single market (1985), the decision to accelerate the liberalization of capital movements (1986), the launching of the project of monetary union (1988), the agreement of Maastricht (1992), and the final adoption of the euro (1998).”

In fact, the Economic and Monetary Union (EMU) was seen by its creators not in purely economic and technical terms, but, in Monnet’s tradition, as “a further step—and as a prerequisite for yet other steps—in the political unification of Europe” (Padoa-Schioppa 2004, p. 6). The same idea was stressed by the first President of the European Central Bank, Wim Duisenberg, according to whom EMU was a “stepping stone on the way to a united Europe.” And this stepping-stone role could be played in spite of (or even as a consequence of) its institutional shortcomings. In the functionalist tradition, the fact that EMU lacked institutions historically associated with a successful monetary union - such as a fully-fledged lender of last resort, a banking union, a fiscal union, and so on - could be rationalized as part of a dynamic path that, in the longer term, would necessarily lead to a political union. For instance, in 1991 German Chancellor Helmut Kohl said: "It is absurd to expect in the long run that you can maintain economic and monetary union without political

---

16 For detailed historical analyses of the negotiations and decisions leading to EMU see Dyson and Featherstone (1999) and James (2012).
In Monnet’s chain-reaction tradition, Kohl’s statement was not meant as a damning assessment of the long-term viability of EMU, but as an optimistic prediction that, eventually, political union would “have to” follow economic and monetary union.

The current crisis in the Euro Area certainly confirms the risks and inefficiencies associated with incomplete institutional integration. Up to a point, recent events have also confirmed that such inefficiencies and crises can create the pressure for more institutional integration – for example, in banking supervision. It is indeed possible (but far from guaranteed) that Europeans will come out of their economic, financial and political crisis with stronger and more deeply integrated institutions. Nonetheless, the crisis has also illustrated the very high costs, risks, and limitations associated with the chain-reaction method of partial integration.

As we have seen, a fundamental problem with the functionalist chain-reaction approach - both as a theory of European integration and as a policy strategy - is that it tends to underestimate the heterogeneity costs and constraints involved when political integration is attempted among populations with different preferences, cultures, and identities.

Stable integration is more likely to take off in areas such as trade, where heterogeneity costs across populations are relatively low, and partly offset by the benefits from diversity. As integration proceeds to other areas, after low-hanging fruits are picked, steeper heterogeneity costs are encountered. At some point, such costs may become politically prohibitive, and stop the process, or even lead to a collapse of the whole system. The risks are particularly high if the previous steps towards more integration have not been taken with the broad democratic consensus of all populations involved.

Consequently, previous success at integrating in areas with lower heterogeneity costs (such as commercial integration) does not necessarily imply further successes at integrating more “political” areas with higher heterogeneity costs (such as fiscal policies or defense). Therefore, the functionalist approach to European integration is really based on a misconception: the expectation that economic integration will lead to political integration. While political unification historically has been used to foster economic integration within a unified domestic market, the opposite does not typically hold. On the contrary, economic integration reached through international cooperation is a substitute rather than a complement

---

19 For a more detailed discussion of these issues see Spolaore (2013, pp. 138-139). An interesting diagnosis along partially similar lines is provided by Mody (2013). For an empirical analysis see Guiso, Sapienza and Zingales (2014), who find that the 1992 Maastricht Treaty, the 2004 enlargement and the 2010 Eurozone crisis seem to have reduced pro-Europe sentiment among European citizens, even though most Europeans still support the common currency. These authors conclude that “Europe seems trapped in catch-22: there is no desire to go backward, no interest in going forward, but it is economically unsustainable to stay still.”
20 For a discussion of the so-called democratic deficit in European institutions see Alesina and Spolaore (2003, chapter 12).
of political integration. If countries can manage to lower barriers to trade among themselves without full political integration, they will face lower incentives to form a political union with a unified domestic market, because such union would generate smaller additional gains from trade. In fact, both theoretical considerations and empirical evidence suggest that international economic integration is associated not with political integration but with political disintegration (Alesina and Spolaore 1997, 2003; Alesina, Spolaore, and Wacziarg 2000).

In sum, the method of gradual and partial integration can lead to stable integration when applied to areas with lower heterogeneity costs and higher economies of scale, but there is no guarantee that it can lead to further integration in areas with much higher heterogeneity costs, or that those costs would endogenously decrease as a consequence of integration.

The formation of a common market, as already mentioned, is an excellent instance of the effective use of partial integration. Overall, the reduction of barriers to economic exchanges was in the general interest of European populations, even though specific sectors within each country benefited from protectionism. As it has often been noted (e.g., Eichengreen 2006), institutional integration in different areas allowed “linkages” between issues and credible side-payments. For example, Europe’s Common Agricultural Policy (CAP) - originally introduced in 1962 and amounting to a substantial share of the European institutions’ budget - can be explained as part of a deal between France and Germany, whereas German taxpayers subsidized French farmers whereas German exporters gained access to the French market.

The creation of a common European legal framework and common supranational institutions has provided national governments with a credible “commitment technology,” going beyond the institutional framework of traditional international organizations. For example, in a landmark case in the early 1960s\textsuperscript{21}, the European Court of Justice decided directly in favor of a Dutch importer of German chemical products that had objected to a tariff charged by the Netherlands in violation of article 12 of the Treaty of Rome. In fact, in this and other cases the European Court of Justice went beyond the legal provisions that had been formally agreed with the Treaty of Rome, and, according to some scholars, brought Europe close to a federal legal system (Weiler 1991, Krasner 1999). These novel legal doctrines, however, were established not in conflict with national governments, but exactly in order to enforce norms consistent with national goals, such as trade liberalization. Therefore, the expansion of powers of the European Court of Justice illustrates the viability of the strategy of supranational institutional integration insofar as it is directed areas with relatively low heterogeneity cost and high economies of scale and externalities. In contrast, supranational

\textsuperscript{21} Van Gend en Loos v Nederlandse Administratie der Belastingen (26/62)
integration and centralization have been met with increasingly binding constraints when attempting to move to more sensitive and political areas. For instance, in recent years Germany’s Constitutional Court has elaborated the legal theory of *conditional* acceptance of the supremacy of European norms, which can be accepted only insofar as they are consistent with “fundamental German rights.” In an important ruling on the Lisbon Treaty in 2009, the German Constitutional Court explicitly called the national states “the masters of the treaties,” and "therefore must see to it that there are no uncontrolled, independent centralization dynamics" within the EU.”

In spite of all its limitations, many supporters of the European project believe that economic integration has benefited Europeans not only directly – through gains from trade – but also indirectly, by reducing the risk of a European conflict. The hypothesis that international trade reduces the risk of war has a long pedigree, going back at least to Montesquieu and Kant, and is part of the broader theory of “liberal peace” brought in by democracy, trade, and international organizations (e.g., Oneal and Russett 1999). Recent empirical studies (Martin, Mayer and Thoenig 2008 and 2010) cast doubt on a positive relation between multilateral openness (globalization) and peace. On the contrary, the ability to trade with third parties reduces the costs of going to war between pairs of countries. *Bilateral* trade, however, by increasing the opportunity cost of war between two countries, lowers the likelihood of conflict between them, even when controlling for the degree of historical, linguistic and religious similarity between their populations (Spolaore and Wacziarg 2013). These studies suggest that regional trade agreements between “old enemies” – such as the formation of a European common market – have probably decreased the risk of conflict among European countries after World War II. An open question is whether European integration has played a major or only a minor role in securing peace in Europe, when compared to other factors, such as the role of the United States and NATO.

5. Conclusions

This paper has discussed the political economy of European integration in light of the implications of the fundamental trade-off between benefits from integration and political costs associated with heterogeneous preferences over public goods and policies.

High heterogeneity costs have so far prevented Europeans from forming a full political union. Attempts to integrate sensitive political functions – such as defense and foreign policy – have not gone very far. Instead, Europeans have adopted a gradual strategy of pooling and delegating functions and policies to supranational institutions in a relatively limited set of

---

22 *BV*erfG, 2 BvE 2/08 of June 6, 2009
23 Quoted in *Spiegel*, 2009.
areas, mostly economic, while maintaining other prerogatives at the national or sub-national level. In spite of supranational rhetoric, ultimate sovereign control and the monopoly of the legitimate use of coercion have firmly remained in the hands of national governments.

This strategy has provided significant economic benefits to Europeans when appropriately implemented in areas with relatively low heterogeneity costs and high economies of scale and scope. However, serious problems and crises have their roots in the expectation that incomplete and partial integration could always be overcome with further integration, in a “chain-reaction” towards an “ever-closer union.” The euro, with its institutional incompleteness and shortcomings, is a child of this strategy.

A realistic political-economy analysis naturally suggests a different, and potentially more effective strategy. The likelihood of achieving durable integration is higher if further steps towards integrating policies and institutions are undertaken only when they are economically beneficial and politically stable on their own merits, and when they are democratically supported by the populations involved.
REFERENCES


Haas, Ernst B. (1975), *The Obsolescence of Regional Integration Theory*, Institute of International Studies, University of California.

Hoffmann, Stanley (1966), “Obsolete or Obstinate? The Fate of the Nation-State and the Case of Western Europe,” *Daedalus*, 95 (3): 862-915.


Moravcsik, Andrew (2012), "Europe After the Crisis," *Foreign Affairs*, May/June


