The Journal of

Economic Perspectives

A journal of the American Economic Association

Celebrating 30 Years

Winter 2016

The Journal of
Economic Perspectives

A journal of the American Economic Association

Editorial offices:
Journal of Economic Perspectives
American Economic Association Publications
2403 Sidney St., #260
Pittsburgh, PA 15203
email: jep@jepjournal.org

The *Journal of Economic Perspectives* gratefully acknowledges the support of Macalester College.
Registered in the US Patent and Trademark Office (®).

Composed by American Economic Association Publications, Pittsburgh, Pennsylvania, USA.
Printed by R. R. Donnelley Company, Jefferson City, Missouri, 65109, USA.

THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Winter 2016, Vol. 30, No. 1. The *JEP* is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. Annual dues for regular membership are $20.00, $30.00, or $40.00 depending on income; for an additional $15.00, you can receive this journal in print. E-reader versions are free. For details and further information on the AEA go to https://www.aeaweb.org/. Periodicals postage paid at Nashville, TN, and at additional mailing offices.
POSTMASTER: Send address changes to the *Journal of Economic Perspectives*, 2014 Broadway, Suite 305, Nashville, TN 37203. Printed in the U.S.A.

The Journal of
Economic Perspectives

Contents

Volume 30 • Number 1 • Winter 2016

Symposia

Articles

Features

Statement of Purpose

The *Journal of Economic Perspectives* attempts to fill a gap between the general interest press and most other academic economics journals. The journal aims to publish articles that will serve several goals: to synthesize and integrate lessons learned from active lines of economic research; to provide economic analysis of public policy issues; to encourage cross-fertilization of ideas among the fields of economics; to offer readers an accessible source for state-of-the-art economic thinking; to suggest directions for future research; to provide insights and readings for classroom use; and to address issues relating to the economics profession. Articles appearing in the journal are normally solicited by the editors and associate editors. Proposals for topics and authors should be directed to the journal office, at the address inside the front cover.

Policy on Data Availability

It is the policy of the *Journal of Economic Perspectives* to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Details of the computations sufficient to permit replication must be provided. The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the above requirements cannot be met.

Policy on Disclosure

Authors of articles appearing in the *Journal of Economic Perspectives* are expected to disclose any potential conflicts of interest that may arise from their consulting activities, financial interests, or other nonacademic activities.

Journal of Economic Perspectives—Volume 30, Number 1—Winter 2016—Pages 3–28

The International Monetary Fund: 70 Years of Reinvention

Carmen M. Reinhart and Christoph Trebesch

As recently as 2008, the International Monetary Fund (IMF) seemed to be winding down its business. After the Argentine and Uruguayan crises of 2001–2003, the world had been comparatively free of financial crises. IMF lending, whether expressed as a share of world GDP or imports, fell to its lowest levels since the early 1970s, as shown in Figure 1. Dollar amounts declined more markedly than the number of programs, as lending to the larger emerging market and middle-income countries mostly came to an end. Loans to low-income countries (involving smaller dollar amounts) became increasingly overrepresented among the remaining IMF programs. A view emerged that perhaps an institution whose primary roles were economic surveillance and crisis management had outlived its usefulness. This interpretation of events may have motivated the IMF to downsize (Obstfeld and Gourinchas 2012). Treating this temporary calm as the "new normal," the IMF shrank the size of its staff, which had expanded considerably in previous decades in response to a sharp increase in its membership (as reported, for example, in *The Economist* 2008).

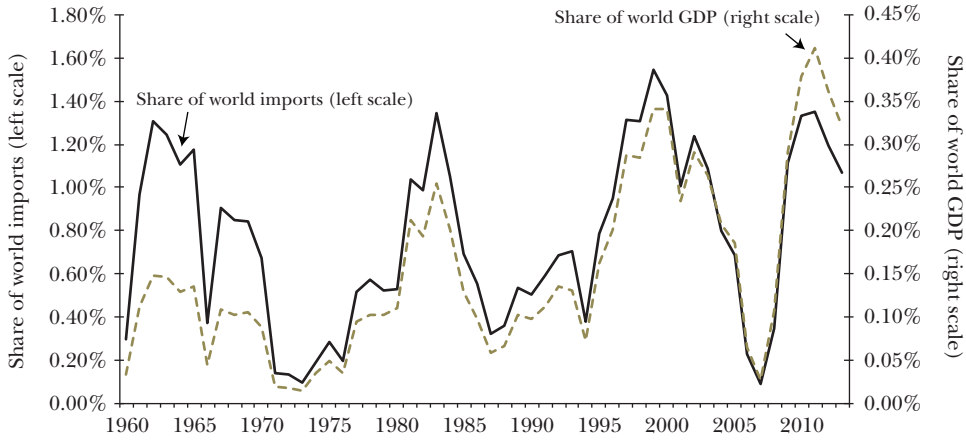
However, the emergence in 2007–2009 of the deepest and most synchronous financial crisis in the world's largest economies since the 1930s put an end to the notion that the IMF was redundant. As Kindleberger (1978) had wisely observed decades earlier, financial crises are "a hardy perennial." By practically any metric,

■ *Carmen M. Reinhart is Minos Zombanakis Professor of the International Financial System, Harvard Kennedy School of Government, Cambridge, Massachusetts. Christoph Trebesch is Assistant Professor, Department of Economics, University of Munich, Germany. Their email addresses are carmen_reinhart@harvard.edu and christoph.trebesch@econ.lmu.de.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.1.3

doi=10.1257/jep.30.1.3

Figure 1
IMF Lending as a Share of World Trade and GDP, 1960–2014



Sources: Gold (1970), International Monetary Fund, *International Financial Statistics*, Monitoring of Fund Arrangements (MONA) Database, *World Economic Outlook*, Joyce (2005), Killick (1995), and Mody and Saravia (2006).

the post-2008 IMF programs to several European economies are the largest in the IMF's 70-year history. As Figure 1 shows, in 2010 the new programs to the wealthier borrowers brought total IMF commitments, measured as a share of imports, close to their historical peak in the early 2000s, while as a share of world GDP, IMF commitments hit an all-time peak.

The IMF has reinvented itself on several occasions and in different dimensions since its creation approximately 70 years ago.[1] Under the Bretton Woods system, the IMF oversaw a network of mutually pegged exchange rates. A key challenge of that system was to get the parity "right." Otherwise, an economy with an overvalued currency would be vulnerable to a weakening in the balance of payments and international reserve losses.

Because exchange rate misalignments and corresponding balance-of-payments problems were a frequent and recurrent preoccupation among the IMF membership, the so-called Stand-By Arrangements of the earlier era involved short-term lending to deal with temporary (illiquidity) problems. This mandate is laid out in section I (v) of the Articles of Agreement of the IMF, which reads: "To give confidence to members by making the general resources of the Fund temporarily available to them under adequate safeguards, thus providing them with opportunity to correct maladjustments in their balance of payments without resorting to measures destructive of national or international prosperity." From this perspective, the IMF was not intended to function as a development agency engaged in long-term lending (this

[1] Excellent companion studies to this paper include Edwards (1989) and Bordo and James (2000). Together, they provide a concise picture of what the IMF does, which is also rich in historical detail.

was a role assigned to the World Bank and various regional development agencies). Nor was it an institution to lend into situations of sovereign insolvency. Its intended mandate was to act as the international lender of last resort.

During the 1950s and 1960s, IMF lending had, indeed, been mostly short-term loans to the governments of advanced economies in connection with comparatively moderate exchange-rate adjustments. But the global Bretton Woods system of pegged exchange rates came apart by the early 1970s, and some of the world's major economies adopted floating exchange rates (for more on "The End of the Bretton Woods System," see the IMF webpage https://www.imf.org/external/about/histend.htm). The role of the IMF began to evolve. At this time, the membership of the IMF also expanded significantly to include a growing number of low-income and middle-income countries. Starting in the late 1970s, the IMF programs increasingly involved lending to countries with a wider range of crises (apart from those related to foreign exchange), including banking and sovereign debt crises. IMF lending in these crises gained ground with the Latin American crisis of the 1980s and the over-borrowing of many transition economies (formerly connected to the Soviet Union) in the 1990s. As banking and debt crises tend to be much more protracted problems compared to the currency crises of the earlier decades, the average duration of IMF involvements increased markedly. Chronic and recurrent IMF clients multiplied, and program duration has sometimes spanned more than 20 years.[2] Thus, IMF programs came to have less to do with the original mission of providing temporary liquidity support, and began to resemble long-term development assistance, especially in some of its poorest member countries.

Since the short-lived lull in the years leading up 2007, the IMF has (once again) redefined its role, making extremely large loans (relative to the size of the national economies) to wealthy economies in Europe, with the largest of these to Greece, where debt sustainability problems have been manifest for some years now. In some sense, this most recent change brings the IMF full circle, because advanced economies had been its earliest and largest clients before the emerging market economies started to dominate its activity in the 1980s.

In what follows, we discuss the evolution of the IMF during the past 70 years from several angles. Our narrative documents the evolving "clientele" for IMF programs and provides a sense of how activity shifted across different parts of the globe and between advanced and emerging economies. We connect some of these shifts to the ascendancy of financial liberalization and the subsequent increase in cross-border finance, as well as global factors, such as international interest rates, primary commodity prices, and global saving patterns. We also consider how the set of situations to which the IMF responds has changed over time, from the early focus on currency problems to the more engulfing challenges posed by systemic banking

[2] Among the studies that have addressed the causes and consequences of the protracted duration of Fund programs, or their recurring "serial" nature, IMF (2002), Bird, Hussain, and Joyce (2004), Joyce (2005), and Mody and Saravia (2006) all figure prominently. Bulow and Rogoff (1999, 2005) discuss alternative approaches to addressing the sustained financing needs of low-income countries.

and sovereign debt crises, often involving protracted output slumps and large-scale bailouts of the corporate banking sectors.

Our approach focuses on the fundamental changes in the IMF's borrowing patterns rather than delving into concerns about how particular crisis episodes were handled. For the latter, there is a substantive body of literature that critiques IMF practices. One strand of this literature focuses on the prominent role of political influence (most notably by the United States) on the design and incidence of IMF programs. Another body of work takes issue with various aspects of IMF conditionality (including the role of fiscal austerity). While we refer to relevant works in these areas, we take a different perspective.

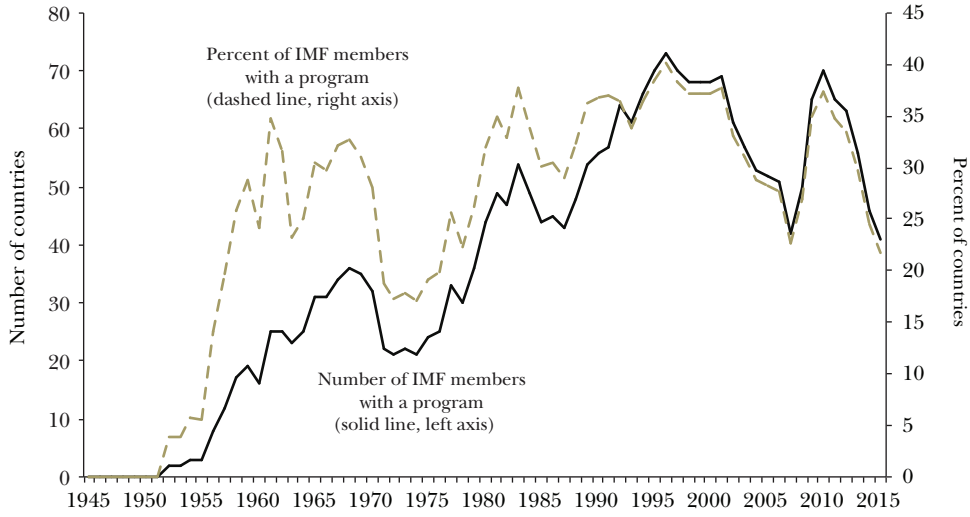
Our main critique focuses on the IMF's role in the international financial architecture and the problems arising from "serial" IMF lending and from lending to countries with excessive debt burdens. We suggest that the changing nature of lending patterns over time has left the IMF with conflicting objectives. In dealing with potentially unsustainable debt cases and in moving toward larger and longer-term loan packages, the institution has become more involved in lending into sovereign default (often chronic). An important unintended consequence of this tilt towards lending into arrears is that a country that seeks an IMF loan may be inadvertently signaling to the rest of the world that it is insolvent (and not just illiquid). Indeed, a recent IMF (2014, p. 4) report recognized that there is a large "stigma associated with using Fund resources." The financial press has been aware of this issue for some time (for example, *The Economist* 2009).

The adverse signaling effect has potentially damaging consequences for the IMF's role as a lender of last resort in crisis times, especially if solvent countries in need of liquidity refrain from approaching the IMF altogether. While there are numerous development banks, the IMF is the one institution that is sometimes described as a central bank for countries or lender of last resort to the world (for example, see Fischer 1999 in this journal, and the essays in Bank of International Settlements 2014). In our view, it is important to draw a clearer line between members that need financing for temporary balance-of-payments or liquidity problems and countries that show a *chronic* dependence on concessionary external funds. At the end of the paper, we offer some speculation on the direction and dimension of the next wave of changes the IMF may confront.

Shifting Clientele

When viewed by the number of its members, the IMF is a very successful institution. The initial membership of 28 back in 1945 has increased steadily to 188 in 2015, with two notable growth spurts: a jump of 33 countries in the early 1960s, as the former colonies in Africa gained independence, and a jump of 28 countries in the early 1990s after breakdown of the Soviet Union (for a list of when countries joined the IMF by date, see https://www.imf.org/external/np/sec/memdir/memdate.htm). New entrants often asked for IMF assistance shortly after becoming

Figure 2
The Incidence of IMF Programs, 1945–2015



Sources: Gold (1970), International Monetary Fund, Monitoring of Fund Arrangements (MONA) Database, Joyce (2005), Killick (1995), and Mody and Saravia (2006).

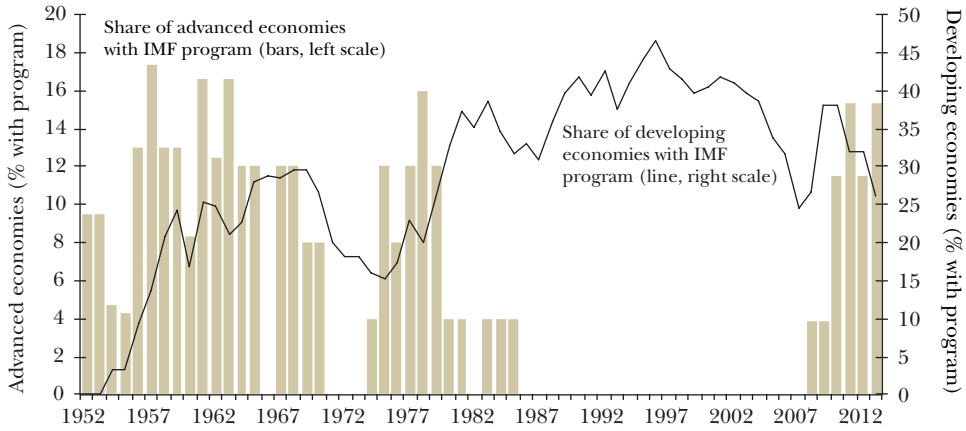
members, which partly explains why the number of countries with an IMF program increased so markedly in the early 1960s, as shown in Figure 2.

The more intense program activity in the developing world also explains why the IMF came to be seen as an institution that uses funds from higher-income countries to grant crisis loans to lower-income countries. Indeed, up until the global financial crisis of 2008–2009, much of the academic and popular debate of the IMF's role in coping with crises was relegated to a discussion of episodes in emerging markets, culminating with Argentina's spectacular default in December 2001. Largely forgotten is the fact that in the decades immediately after its creation at the end of World War II, many IMF programs were providing balance-of-payments support to the comparatively wealthy economies of Europe. Figure 3 documents the incidence of IMF programs over 1950–2014 in advanced and emerging market/developing country groups separately, making plain the swings in the pendulum from advanced economies in the 1950s and 1960s, to emerging markets in the mid-1980s, and back to advanced economies after a 30-year hiatus in 2008.

Early Decades: Significant Lending to Advanced Economies

There were no IMF lending programs in the first seven years after its creation in 1945. Rather, the Marshall Plan of 1948–1951 was the dominant form of international transfers (via both loans and grants) to 18 European countries (if the Free Territory of Trieste is counted), along with 8 Asian countries, as well as Israel and a few other countries in the Middle East. While the Marshall plan is usually discussed

Figure 3
From Advanced Economies to Emerging Market (Developing) Economies and Back: Program Incidence, 1950–2014



Sources: Gold (1970), International Monetary Fund, Monitoring of Fund Arrangements (MONA) Database, Joyce (2005), Killick (1995), and Mody and Saravia (2006).
Note: We use the terms "emerging market economies" and "developing economies" interchangeably, here and throughout the paper.

in terms of humanitarian aid and reconstruction finance, its loans and grants also provided much-needed balance of payments support, as the economies in question faced the twin challenges of having significant needs for imported consumer and industrial goods and limited reserves of hard currency to make such purchases. This problem was known at the time as the "dollar gap" (for discussion, see Behrman 2007, especially chap. 8). Some countries, most notably the United Kingdom, had also inherited a high level of wartime debt to service. In others, including France, Germany, Italy, and Japan, roaring inflation at the end of the war and the beginning of the peace had wiped out significant portions of the domestic stock of debt (Reinhart and Sbrancia 2015).

In 1952, Belgium and Finland were the first countries to use IMF resources. Between 1956 and 1977, the United Kingdom alone had 11 IMF programs. In the early years of the IMF, the classic lending instrument was a Stand-By Arrangement, in which funds are provided on the condition that the borrower addresses its underlying imbalances and were typically granted as one-year programs that in a few cases were renewed. By 1977, at the outset of the last of the UK programs, the duration of that Stand-By Arrangement was stretched to two years. This trend has persisted. For example, Turkey's most recent Stand-By program, in 1999, was a three-year arrangement, reflecting a reinterpretation of what constitutes temporary support (an issue we take up in the next section). The current structure for Stand-By Arrangements, updated most recently in 2009, is explained at https://www.imf.org/external/np/exr/facts/sba.htm.

In these early decades of the IMF, France, Iceland, Italy, Portugal, Spain, and the United States, among others, all borrowed from the International Monetary Fund. The fact that Greece did not have an IMF program during this era largely stems from the fact that the country was in default from 1932 through 1964 and had rather minimal interaction with world capital markets. It was also the case that Greece was a major recipient of aid, rather than loans, during the encompassing umbrella of the Marshall Plan. The composition of the early developing-country clients of Fund also differed from what was to emerge in the decades that followed. In the pre-OPEC era (OPEC was founded in 1961), Iran and Syria had very brief stints with Fund programs. This has not been repeated since. South Africa was the only African country in the IMF until the late 1950s, when Ghana and Morocco joined the membership. However, some of the countries that were to become chronically attached to the IMF (as we discuss in the next section)—Argentina, El Salvador, Pakistan, the Philippines, among others—had already made their appearance at the IMF lending window by the late 1950s or early 1960s.

Emerging Market, Transition Economies, and Turmoil

Many emerging market economies experienced an economic boom in the 1970s. For some of these countries, the driving force was the sharp rises in oil prices in the mid and late 1970s (Díaz-Alejandro 1983, 1984). The surge in oil prices, accompanied by worldwide inflation, lifted commodity prices in general. Believing these higher commodity prices would last into the long-term, many came to view lending to commodity producers as a lucrative activity. A common dynamic of international finance at this time was the so-called "recycling of the petrodollars," which refers to the recurring pattern whereby oil-producing countries deposited their surging dollar-denominated export revenues in international financial intermediaries that, in turn, aggressively expanded their lending to a broad range of developing countries.

The spectacular boom was followed by a protracted bust—a common historical pattern (Kindelberger 1978; Kaminsky and Reinhart 1999; Reinhart and Rogoff 2009; Gourinchas and Obstfeld 2012). The spike in US interest rates in October 1979 abruptly brought the feast-phase of the cycle to an end. Given that a significant share of the new debts of emerging market economies were either short term or carried a variable interest rate, there was a swift and adverse effect on their national balance sheets. Furthermore, the sharp appreciation of the US dollar accompanying the Federal Reserve's tight monetary policy further undermined the solvency of those nations that had taken on dollar-denominated debt (whether those debts were public or private). By the early 1980s, commodity and oil prices plunged, and debt-servicing costs for the developing world skyrocketed. The decade of the 1980s is often referred to as the "lost decade" for Latin America. While emerging markets in Asia fared better in terms of growth, inflation, and gains in social indicators, commodity-intensive Africa fared just as badly in general and in several dimensions worse. For emerging markets as a whole, the 1980s was the most dismal decade since the 1930s.

The number of IMF lending programs more than doubled from 1976 to 1983. The new wave of IMF clients was comprised mostly of what in the language of the day were called "less-developed countries" or LDCs. This marked shift in the composition of types of countries borrowing from the IMF influenced the modalities and scope of the programs. In 1987, the IMF introduced the Enhanced Structural Adjustment Facility (ESAF) program, which focused on making low-interest loans to low-income countries. The debt crisis of the 1980s was eventually addressed after several years. One main step was the external debt restructurings in 16 countries (mostly in Latin America) under the Brady Plan of 1989, in which creditors agreed to write down the debts they were owed in exchange for being issued new debt that was more likely to be repaid. The IMF participated by setting aside some of its own loans. In addition, high-inflation countries in Latin America and elsewhere undertook macroeconomic stabilization and anti-inflation programs.

In the 1990s, the breakdown of the Soviet Union and its satellite community ushered new clients to the IMF. A rough categorization of the new members would place much of Eastern Europe in the middle- to high-income category and the former Soviet republics in the lower-income strata. From this lower-income group, in particular, a number of countries have become recurring and chronic users of Fund resources since joining the Fund in the 1991–1993 period. As the next section documents, this further lengthened the "effective" duration of IMF programs.

Post-2008: Crisis in the Eurozone

In the years before the global financial crisis erupted in 2008, many countries in the periphery of Europe, as well as Iceland, the United Kingdom, and the United States, experienced a boom in capital inflows. As in so many pre-crisis booms, borrowing from the rest of the world supported a combination of larger current account deficits, domestic lending surges, and asset price booms. For example, in 2008, the current account deficits in Iceland, Greece, Portugal, and Ireland had reached records of 28, 15, 13, and 6 percent of GDP, respectively. As the global financial crisis severely sapped economic activity and confidence in sovereign government finances eroded, the access to international capital markets that had been taken for granted by advanced economies during most of the post–World War II era came to a sudden stop. Iceland was the first to start an IMF program in 2008, followed by Greece and Ireland in 2010 and Portugal in 2011.

This episode brought an enormous rise in the volume of lending (in real terms) directed at the higher-income economies: from 2008 to the present, the IMF has loaned more than $200 billion, of which about two-thirds went to advanced economies like Greece, Iceland, Ireland, and Portugal. This emphasis on lending to advanced countries represented the reemergence of an earlier pattern. In the 1950s and 1960s, advanced economies accounted for a larger share of total approved lending than emerging and developing economies. Table 1 shows the largest IMF programs during its history, as measured by the dollar amounts of the loans approved (in real 2009 dollars) and in relation to the country's GDP and quota. The quota is a subscription fee that must be paid by the country to the

Table 1
The Largest IMF Programs

Country, program year	*Billions 2009 US$*	*Percent of quota*	*Percent of GDP*
Greece 2010	39.851	3,212	13.71
Ireland 2010	29.347	2,322	14.19
Portugal 2011	36.326	2,306	15.76
Greece 2012	34.700	2,159	14.67
South Korea 1997	27.309	1,938	3.81
Turkey 1999	25.674	1,560	8.23
Turkey 2002	19.519	1,330	7.14
Romania 2009	17.645	1,111	10.74
Hungary 2008	16.783	1,015	10.80
Brazil 2002	41.677	902	7.03
Ukraine 2008	17.518	802	9.66
Argentina 2000	27.280	800	6.49
Ukraine 2010	15.076	729	11.19
Pakistan 2008	11.524	700	6.72
Turkey 2005	10.697	691	2.03
Mexico 1995	24.284	688	5.33
Indonesia 1997	14.690	557	5.32
Brazil 1998	17.912	480	1.68
Argentina 2003	14.504	424	8.01
Brazil 2001	18.454	400	2.79
Russian Federation 1996	24.976	306	4.90
India 1981	12.115	291	2.99
United Kingdom 1977	11.146	120	1.53

Sources: Gold (1970), International Monetary Fund, Monitoring of Fund Arrangements (MONA) Database, Joyce (2005), Killick (1995), and Mody and Saravia (2006).

IMF upon joining. It used to be that a country could expect to be able to borrow in the range of 200 to 600 percent of its quota, though this range has been greatly exceeded in recent years.[3]

The Mexican and Indonesian programs of 1995 and 1997 made headlines at the time with loans of 6 to 7 times their quota but they are dwarfed by the more recent wave of IMF lending. The sheer scale of IMF lending during the most recent crisis, whether the country is in the advanced or emerging category (like Hungary, Romania, and Ukraine), is a multiple of what were considered record-sized programs in the 1990s and early 2000s.

[3] The total of all national quotas is at present around $330 billion (although the IMF denominates the total in "Special Drawing Rights" whose value changes with movements in exchange rates). The main factor determining the quota is the size of a country's GDP, but a country's openness to trade, economic variability, and size of international reserves may also play a role.

If one scales IMF lending by the recipient country's GDP, the median program was less than 2 percent of GDP through the late 1970s—which is consistent with what might be expected in connection with temporary balance-of-payments support. For several years during the various debt crises of the 1980s, the median IMF program reached about 4 percent of GDP and remained in that range through the world-wide lull in financial crises during 2003–2007. The sheer size of the post-2008 IMF programs has no historical antecedent, as programs accounted for 10–16 percent of the GDP of recipients. In addition, these post-2008 programs offered the same three-year duration of loans that the Extended Credit Facility offered to low-income countries.

Evolving Demands

As the composition of the IMF's clientele evolved during the past 70 years since the institution's birth, so did common challenges or "types" of crises facing the IMF and its membership.

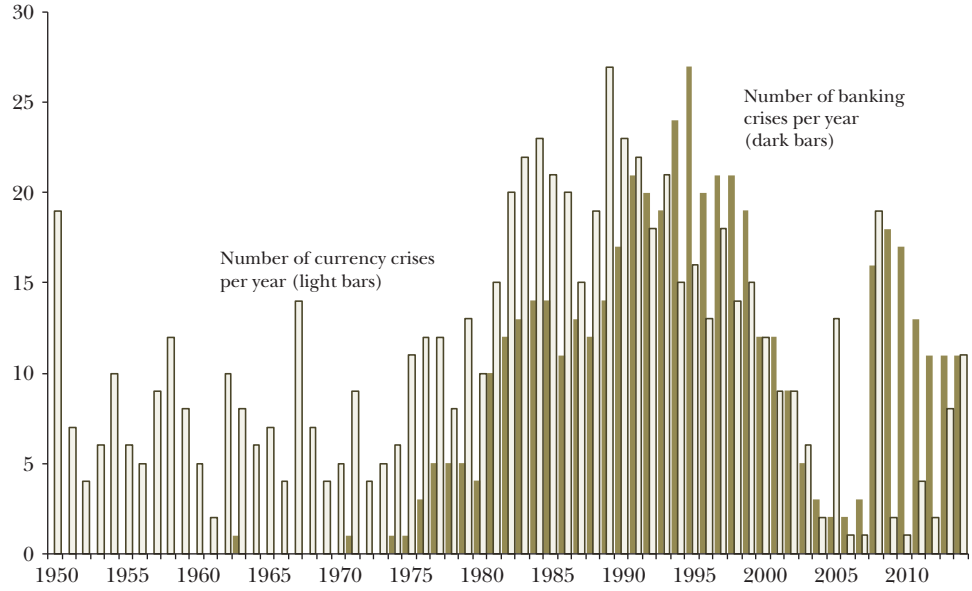
From Currency Crises to Banking and Sovereign Debt Crises

Currency crises (or realignments) were not uncommon during the Bretton Woods era, as highlighted by the light bars in Figure 4 plotting the number of currency crises per year over 1950–2014. The year-by-year count of banking crises is given by the dark bars. As the figure makes plain, while realignments (often in the form of large devaluations followed by re-pegging) were commonplace, banking crises during the era of capital controls and tightly regulated financial markets were rare.[4] Owing to a combination of a broader adoption of more flexible exchange rate arrangements and a decade of prosperity in emerging markets, the incidence of currency crashes diminished through 2012. As emerging-market exchange-rate volatility has climbed noticeably (once again) it remains to be seen whether currency woes resurface.

Most currency realignments are essentially temporary disturbances. Cooper (1971) and Krueger (1978), for example, reach this conclusion after reviewing large currency devaluation episodes during 1951–1970 (many of which were a part of the conditionality associated with an IMF adjustment program). Kaminsky and Reinhart (1999) compare the effects of currency crises and banking crises, and "twin" crises involving both, in a more up-to-date sample. In all of these studies, currency devaluations were associated with either a decline in output or a slowdown in growth, but the effects were comparatively small and short-lived. In contrast, it is a well-documented regularity that recessions associated with systemic banking crises tend to be severe and protracted, as the global experience since 2008 has

[4] Largely owing to a combination of a broader adoption of more flexible exchange rate arrangements and a decade of prosperity in emerging markets during 2003–2013, the incidence of currency crashes in recent years has diminished in comparison to the 1950s–1970s (let alone the turbulent 1980s).

Figure 4
Number of Banking and Currency Crises per Year, 1950–2014



Sources: Reinhart and Rogoff (2009) and sources cited therein; updates by Reinhart at http://www.carmenreinhart.com.

illustrated anew (for broad coverage of this issue, see Kaminsky and Reinhart 1999; Dell'Ariccia, Detragiache, and Rajan 2008; Reinhart and Rogoff 2009; Claessens, Kose, and Terrones 2009; Reinhart and Reinhart 2010; Gourinchas and Obstfeld 2012; and Jordà, Schularick, and Taylor 2013). Cases of sovereign insolvency often involve even more protracted slumps.

Taken together, these observations about the relative incidence of currency versus banking crises, and their comparative speeds of recovery, has implications for the kind of support the IMF may have deemed appropriate.

Growing Duration of IMF Lending Programs

During the 1950s and 1960s, according to our calculations, the duration of IMF lending programs oscillated in a one- to two-year range (the Stand-By Arrangements during the 1950s and 1960s were one-year programs, but a succession of one-year programs was possible). By the end of the 1990s, average duration of an IMF interaction had climbed to three years.

The presence of frequent systemic banking crises may partially account for a lengthening in the duration of IMF programs in 1980–1990s and for an increase in their size relative to GDP. But sovereign default is more protracted still; indeed, sovereign default spells lasting a decade are not uncommon (Reinhart and Rogoff 2009; Cruces and Trebesch 2013).

Table 2
Share of Years with an IMF Program since Joining the Fund

Frequency distribution	*Number of countries*	*Share of countries*
No IMF programs	42	22.3%
In a program 0–10% of the years	23	12.2%
In a program 10–20% of the years	18	9.6%
In a program 20–30% of the years	16	8.5%
In a program 30–40% of the years	19	10.1%
In a program 40–50% of the years	22	11.7%
In a program $\geq$ 50% of the years	48	25.5%
Total	**188**	**100%**

Sources: Gold (1970), International Monetary Fund, Monitoring of Fund Arrangements (MONA) Database, Joyce (2005), Killick (1995), Mody and Saravia (2006), and authors' calculations.

We also find an increasing number of repeated IMF programs, resulting in "serial lending." Table 2 presents a frequency distribution for share of years that a country had an IMF program during the time the country belonged to the IMF. Of the current 188 member countries, more than a quarter of these (25.5 percent) have had an IMF program 50 percent of the time (or more) that they were an IMF member; 37 percent of the countries have been on IMF programs 40 percent of the time or more. Forty-two countries (22 percent) never had a Fund program, with oil-exporting countries and small states accounting for a significant share of this latter group.

Table 3 lists countries with the most recurring use of IMF resources, showing both the share of years with IMF programs (breaking out the data summarized for all member countries in Table 2) and the longest spell (in years) of consecutive programs. Topping the list, Uganda and Malawi (both member countries since the mid-1960s) have had consecutive IMF programs for nearly 30 years. The Joyce (2005) study suggested income levels were an important factor in explaining the duration of programs. With lower-income countries having limited or no access to private capital markets during long stretches of time, one potential story is that the IMF (along with other sources of official financing) has emerged as a near-permanent substitute for access to private capital markets for many low-income countries. It is unlikely, however, that this is the whole story because the list of countries with serial IMF programs shown in Table 3 includes a substantial number of middle-income countries using IMF programs in over 30 percent of the years since their membership.

Chronic Debt Burdens and Lending into Insolvency

The heavy use of IMF lending by so many countries is clearly not a result of short-term currency crises, and the historical incidence of banking crises does not seem sufficient to explain it either (Reinhart and Rogoff 2009, Ch. 10). More

Table 3

Countries with Heaviest Recurring Use of IMF Programs: Incidence and Durations of Spells

Country	*Share of years with programs*	*Longest spell (years)*	*Membership year*
Uganda	67.9%	29	1963
Malawi	70.6%	28	1965
Burkina Faso	47.2%	25	1963
Argentina	60.0%	24	1956
Mali	71.7%	21	1963
Haiti	71.4%	21	1953
Mauritania	62.3%	20	1963
Tanzania	63.0%	20	1962
Togo	44.4%	20	1962
Philippines	53.5%	20	1945
Jamaica	56.6%	20	1963
Panama	47.1%	20	1946
Guinea	54.7%	19	1963
Colombia	46.5%	18	1945
Costa Rica	41.4%	18	1946
Peru	60.6%	18	1945
Gabon	52.8%	17	1963
Zambia	60.8%	17	1965
Bulgaria	65.4%	17	1990
Guyana	68.0%	17	1966
Mozambique	87.5%	16	1984
Romania	70.5%	16	1972
Bolivia	54.9%	16	1945
El Salvador	51.4%	16	1946
Jordan	31.3%	16	1952
Benin	47.2%	15	1963
DR Congo	47.2%	15	1963
Liberia	55.6%	15	1962
Sierra Leone	68.5%	15	1962
Nicaragua	50.0%	15	1946
Armenia	83.3%	15	1992
Georgia	83.3%	15	1992
Madagascar	54.7%	14	1963
Morocco	48.3%	14	1958
Senegal	61.1%	14	1962
Mongolia	64.0%	14	1991
Algeria	24.5%	13	1963
South Korea	41.0%	13	1955
Paraguay	26.8%	13	1945
Uruguay	58.6%	13	1946
Burundi	52.8%	12	1963
Cote d'Ivoire	58.5%	12	1963
Ghana	55.9%	12	1957
Kenya	67.3%	12	1964
Albania	80.0%	12	1991
Honduras	63.4%	12	1945

Notes: The share of years with programs is calculated from the year the country becomes a member (shown in the last column) through 2015.

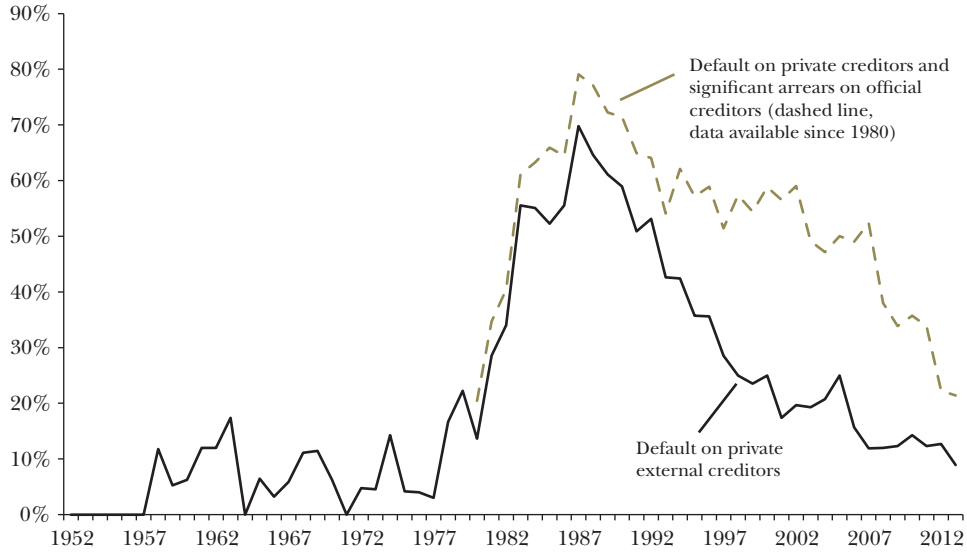
plausible is that IMF lending programs increasingly occur in countries facing problems with insolvency of sovereign and sometimes private debt. A starting point in examining the nexus between IMF program lending and debt defaults is to determine the extent of overlap between the two. The chronology of sovereign default and restructuring on external debt in Reinhart and Rogoff (2009), which we update in Reinhart and Trebesch (forthcoming), covers sovereign default and restructuring on private creditors. These highly visible credit events involve bank loans, bonds, or both. In numerous episodes, this chronology has also identified significant defaults on trade credit as well (see also the analysis by Erce 2013).

As we note in Reinhart and Trebesch (forthcoming), a fuller picture of solvency also requires an assessment of a debtor country's standing with its official creditors. Indeed, the most prominent debt crisis of the last few years, the situation in Greece, now revolves almost entirely around the country's debts to official creditors including the IMF. While official creditors are not the main story for most middle-to-high income countries, they play a dominant role in many low-income countries. It is important, therefore to also assess to what extent *official* debt is in default, under restructuring, or in substantial arrears. This task was recently attempted by Beers and Nadeau (2015), based mainly on World Bank and Paris Club data on defaults and arrears with official creditors. We use their data to complement our earlier history of private and sovereign credit events, which allows us to study the overlap between the augmented (private plus official) defaults and IMF programs on a country-by-country basis.

The pattern we observe is that the share of IMF programs with countries that are either in default or in the process of restructuring with private creditors climbs from less than 20 percent of all programs during the 1950s–1970s to around 70 percent in the late 1980s. The first country to have an IMF program overlap with a default was Argentina in 1958 during the post-Peron budget crisis. The more famous defaults began in earnest in the summer of 1982, as Mexico defaulted in August of that year. However, smaller countries like Bolivia and much of Central America and numerous African countries were already in default as early as 1980. During the 1990s, however, the share of IMF programs involving private debt default then started to fall. Part of the reason was the aforementioned Brady plan in the early 1990s, which diminished the number of debtors in default. Another important step in reducing debt burdens involved the Heavily Indebted Poor Countries (HIPC) Initiative started by the IMF and the World Bank in 1996, which eventually allowed 39 low-income countries to have existing debts reduced in exchange for reforms that made paying the remainder of the debt more likely. (For more information on the HIPC Initiative, see this IMF FactSheet dated September 17, 2015: https://www.imf.org/external/np/exr/facts/hipc.htm.)

Many emerging markets also experienced widespread public and private deleveraging following the Asian crisis of 1997–1998, which made debt defaults in these countries less likely. Finally, the increase in IMF program size, coupled with large-scale bailouts by other official lenders (like the US Treasury or Eurozone institutions like the European Financial Stability Facility and the European Stability

Figure 5
Share of IMF Programs that Overlap with a Default or Restructuring Spell: 1952–2013



Sources and Notes: Data on sovereign defaults to private external creditors is from Reinhart and Rogoff (2009), Cruces and Trebesch (2013), and Reinhart and Trebesch (forthcoming). The dotted line adds to this instances of default on official creditors and their overlap with IMF programs, with data available since 1980. A country is coded as having "significant and persistent arrears" to official creditors if these arrears (including to the IMF and World Bank) exceed 1 percent of GDP for three consecutive years or more, using data on official arrears from the World Bank (2015) and Beers and Nadeau (2015). Data on GDP in current US dollars is from the IMF *World Economic Outlook* database.

Mechanism) have made sovereign default less likely, as private creditors are repaid with the new official loans. By 2014, fewer than 10 percent of all IMF lending programs involved a default to private creditors, as shown in Figure 5.

Figure 5 also demonstrates that the private-only measure of default shows just part of the picture and significantly understates the weak and chronic state of fiscal finances, in particular in low-income countries (and since 2010 in several European countries). Credit events involving official creditors do not often make headlines, with the notable exception of Greece in 2015 (even the restructuring of significant levels of official debt in Ireland, Greece, and Portugal in 2013 received little attention). Despite the limited coverage, arrears on official debts are both large and frequent, as shown in Schlegl, Trebesch, and Wright (2015). Accordingly, the dotted line in Figure 5 shows that defaults on official debt are a frequent by-product of IMF lending programs until this day.

More specifically, in the 1990s and 2000s, about 40 percent of all IMF lending programs involve some sort of default, restructuring, or arrears on official debt. This is a remarkable and not widely known fact, and is surprising given that the

Fund's policies dictate (at least in principle) that the Fund should not lend in the presence of arrears to official creditors.[5] Under exceptional circumstances this rule of nontoleration of arrears can be waived, for example when countries are in the process of negotiating over debt relief with official creditors. However, the data suggests that arrears to official creditors are the norm rather than the exception in many poor countries borrowing from the IMF.

The intersection between lending programs and sovereign defaults or arrears provides insight as to why so many countries have a track record that is filled with year upon year of consecutive IMF lending programs. Indeed, one of the reasons why some countries have become chronic borrowers from the IMF is that it is an effective way of ever-greening their ongoing loans to both private and official creditors. Another reason is that countries asking for debt relief from the Paris Club or under the HIPC Initiative are often asked to agree to an IMF adjustment program first (Rose 2005 finds that 80 percent of Paris Club deals coincide with an IMF program signed in the same year). Yet we do not suggest that debt problems are a complete explanation for the serial lending patterns we observe. Stone (2004) provides compelling evidence for 53 African countries from 1990 to 2000 which shows that the typical IMF's loans-for-reform contract lacks credibility because donor countries intervene to prevent rigorous enforcement; specifically, countries with influential developed-country patrons are subject to less-rigorous enforcement (as measured in terms of shorter program suspensions). Bird, Hussain, and Joyce (2004), who focus on IMF lending programs over 1980–1996, find that recidivist borrowers have lower reserve holdings, larger current account deficits and capital outflows, lower but less-volatile terms of trade, larger debt service and external debt ratios, lower investment rates and per-capita income, and weak governance. Barro and Lee (2005) conclude that both economic factors and measures of political and economic connections to the United States and, to a lesser degree, to the major European countries play significant roles in raising the probability and size of IMF lending. Collectively, these studies suggest that both economic fundamentals as well as political influence help explain why so many countries are "addicted to the IMF."

Implications of the Eurozone Crisis

The euro-zone crisis was the latest demand "shock" for IMF lending. It had a major impact on the IMF's loan portfolio and the IMF's perception as a lender of last resort (Moreno 2013; Schadler 2013, 2014).[6] Three features of IMF involvement stand out.

First, the crisis in Europe strengthened the tendency towards bigger programs and towards lending to countries with very high levels of debt. This increased the

[5] Most recently, in the context of the Ukraine debt restructuring, an IMF spokesman confirmed: "The Fund does maintain a policy of non-toleration of arrears to official bilateral creditors." See Transcript of a Press Briefing by William Murray, Deputy Spokesperson, https://www.imf.org/external/np/tr/2015/tr032615.htm.

[6] The Online Appendix available with this paper, specifically Figures A.2, A.3, and A.4, illustrates the points made in this section.

risks incurred by the IMF considerably. Table 1 shows that the recent programs in Greece, Ireland, and Portugal beat records in terms of size in total amounts (and also when expressed as a percent of debtor country GDP). All three countries have a debt/GDP ratio above 100 percent. Moreover, since 2008, several other European countries received major IMF loans, in particular Iceland, Ukraine, and Hungary. As a result, the IMF loans portfolio became highly concentrated in Europe. As of 2013, almost 80 percent of outstanding IMF loans were owed by European countries alone. The scale of concentration is unprecedented in the IMF's history.

Looking back, we know that some of the European IMF programs did not fail, in the sense that the IMF credits were largely paid back in Iceland, Ireland, and Portugal. In the cases of Greece and Ukraine, which remain large-scale debtors, the jury is still out. However, it seems obvious that, in terms of its loans portfolio, the risks incurred by the IMF after 2008 were larger than ever before.

A second feature of the European crisis is the damage it did to the IMF's reputation, in particular, in the case of Greece. Most visibly, the Greek default on IMF loans was a blow to the IMF's seniority status. When Greece missed a payment of €1.5 billion on June 30, 2015, and another payment of €456 million on July 13, 2015, it became the first advanced economy ever to miss an IMF payment. In the past 70 years, 23 countries had run into protracted arrears with the IMF, but the large majority of these defaulters were low-income countries or countries suffering from severe war or natural disasters. Defaulting on the IMF is typically a last recourse. As shown by Schlegl, Trebesch, and Wright (2015), the Fund is at the top of the "pecking order" of sovereign debt repayments, as debtor countries usually default on all other external creditors first, prior to missing any payments towards the IMF. Notably, this was not the case in Greece, which defaulted on the IMF but continued to repay private bondholders. At the same time, Greece was by far the largest client of the IMF, accounting for 26.6 percent of total IMF lending in end-2014.

Ultimately, a full-fledged default was avoided because the Eurozone governments agreed to new loans, which Greece used to repay its IMF debt. But the most recent installment of the Greek crisis (which is far form over) shows just how dangerous it is for a lender of last resort to agree to serial lending to a country with unsustainable debts. Lending into insolvency (especially in large member countries) endangers the IMF's most valuable asset: its seniority. This is especially true because the IMF's seniority is not written in law, but rather is a market convention. If market participants and debtor governments start questioning the IMF's seniority status, there is little the IMF can do to enforce this status.

More generally, the IMF's role in the Greek crisis has been widely criticized (Subramanian 2015; Mody 2015), and the Fund has publicly acknowledged mistakes (IMF 2013). In a longer historical perspective, however, this dimension is hardly new, as the institution has come under repeated fire in its handling of past crises like the ones in Latin America during the 1980s, in East Asia from 1997–1998, in Russia in 1998, and in Argentina in 2001, among others.

The third and maybe most problematic legacy of the Eurozone crisis is the so called "systemic exemption" clause (IMF 2014b), which effectively scrapped the IMF's

long-standing rule that no loans should be given to countries with unsustainable debts. The policy authorizes the IMF to lend to any country (even insolvent ones) if this country poses large risks of systemic financial spillovers. It was introduced in 2010, when the Greek debt burden was no longer evaluated as "sustainable with high probability" by Fund staff (a precondition for grating exceptionally large IMF loans). In response, the IMF altered its lending framework ad hoc, argued that Greece indeed posed severe spillover risks for the Eurozone and the world economy, and, on the same day, granted the country access to IMF loans worth €30 billion—more than 3000 percent of the Greek quota (the largest Fund program ever relative to quota, see IMF 2013). Despite a recent staff proposal to drop the exemption clause, the rule remains in place until this day, meaning that IMF lending into insolvency can continue on a large scale, at least as long as IMF management judges the member country to be of systemic importance. As a result, the IMF may face larger demands for new loans, including by large emerging markets with heavy debt burdens. Moreover, credibility problems may increase in cases where private creditors are asked to agree to a debt restructuring and share part of the burden in future bailout programs. Such developments may contribute to creditor moral hazard and further undermine the Fund's role as a lender of last resort.

Demand-Driven or Supply-Driven?

Discussions of the trends and patterns in IMF program lending often fall into two main categories depending on whether they tend to emphasize demand-driven or supply-driven explanations. In explaining the breadth and timing of the changes in IMF lending documented in this paper, we lean toward a "borrower demand–driven" theory of institutional change at the IMF, in which the Fund has redefined the issues it seeks to address and the tools it employs based on the evolving needs of its clientele.

One important demand factor of this sort has been the greater re-globalization of capital markets.[7] Banking crises had been rare for several decades prior to the re-globalization of capital markets in the early 1980s. The increasing frequency of banking problems in IMF member countries, and in particular the experience of the Mexican and East Asian crisis of the 1990s, paved the way for the introduction of a new department within the IMF that focused on the financial sector. Moreover, by the late 1990s, the Financial Sector Assessment Program (FSAP) had become a central element of IMF "surveillance," in which it evaluates the economic risks that countries may face. Another demand factor has been the persistent financing need of poor countries without access to private external capital markets, which in 1987, gave rise to the Enhanced Structural Adjustment Facility (ESAF) program focused on low-interest loans to low-income countries. As noted, a significant share of these

[7] "Re-globalization" refers to the fact that the integration of capital markets was substantial in the heyday of the gold standard era (Eichengreen 1992; Obstfeld and Taylor 2004; Reinhart and Rogoff 2009), before two World Wars, with the global economic depression of the 1930s in between, balkanized global finance.

programs involve countries with chronic arrears to official borrowers and account for some of the longest (recurring) IMF programs.

An alternative view would characterize the IMF as being "creditor supply–driven," taking the position that key funders (and the United States in particular) have dictated to the IMF and used it as an extension of their own treasury ministries. This argument is neither new, nor difficult to understand. The voting structure of the IMF is based on the size of "quotas" or financial resources devoted to the Fund, which are in turn linked to the size of national economies, so the United States and major economies of Western Europe have IMF quotas that have traditionally granted them considerable power over IMF decisions. (For instance, the United States has a large enough share of total votes that it can exercise veto power over any substantial IMF decision, while the Managing Director of the IMF has always been a European.) Indeed, there is a compelling range of evidence to support the conclusion that politics plays a role in IMF lending decisions (usually at the country level). We have already alluded to some of these findings (for example, the Barro and Lee 2005 study). There are other studies. Thacker (1999), for instance, finds that political factors and voting alignments with the United States are significant in explaining the probability of getting an IMF loan (although his results vary across sample periods). Stone (2004) discusses the political economy of IMF loans in Asia. Dreher, Sturm, and Vreeland (2009) find systematic evidence that UN Security Council membership reduces the number of conditions included in IMF programs. Another recurring critique is that the use and character of the conditions that the IMF places on its loans (as well as the rigor with which such conditions are enforced) may be unduly influenced by the objectives of key lenders. Jeffrey Sachs (1998) was a vocal proponent of this view at the time of the Asian crisis from 1997–1998, while Feldstein (1998) is explicit about this problem in the case of Korea. More systematic evidence comes from Dreher and Jensen (2007), suggesting that closer US allies face less-strict IMF conditions.

More broadly, the issue of political control over the IMF has been evolving. For example, China's current quota is less than one-quarter of that of the United States even though by 2014 its share in world GDP (adjusted for purchasing power differentials, as reported by the IMF's *World Economic Outlook*) reached approximately the same as that of the United States. The China–US split on this issue is much broader (and much older), because it represents the tension between the advanced and developing country membership within the IMF (which was expressed as North versus South in an earlier literature). IMF quotas, with their implications for voting power within the institution and the quantity of loans that can be received, are reviewed infrequently. There has been a proposal on the table since 2010 for resetting the IMF quotas. The proposal would roughly double the size of the quotas—and thus double the lending power of the IMF—but it would also diminish the relative power of the United States and Western Europe. China would become the country with the third-largest quota and voting power at the IMF, and Brazil, India, and Russia would all be in the ten largest countries by voting power (for more details on the proposed change, see this

IMF FactSheet on quotas, dated September 24, 2015: http://www.imf.org/external/np/exr/facts/quotas.htm). However, the change cannot occur without the United States supporting the proposal, which it has so far declined to do.

In sum, while we would emphasize a "demand driven" interpretation for general or aggregate trends in IMF lending, our view does not preclude the possibility that political factors play a significant role in the design or implementation of IMF lending programs for individual countries, as suggested by much of the evidence.

Conflicting Objectives

During the decades since the IMF was founded in 1945, a clear disconnect has emerged between the institution's original mandate and its modern operations. Indeed, the gap between mandate and operations appears to have widened over time. The original mandate of the IMF focused on temporary lending when liquidity was tight and balance of payments support was needed. Despite those intentions, we have documented here a number of patterns that suggest an alternative mission has taken root: 1) about one-quarter of the member countries have been engaged in an IMF program more than half of the years since becoming a member; 2) 66 out of 188 member countries have had consecutive IMF program spells that last somewhere between 10 and 30 years; 3) since the 1980s, the share of IMF programs involving a sovereign that is in default, restructuring, or arrears has oscillated between 40 and 70 percent; and 4) IMF lending in the post-2008 period to high-income countries in Europe has reinforced the prior ongoing trends toward larger and longer programs that are entangled with issues of sovereign debt.

In explaining the reasons behind this change, we lean toward an interpretation in which the Fund has redefined the issues it seeks to address and the tools it employs based on the evolving challenges of its members. A more negative interpretation is that the trends documented here are evidence of mission creep at the IMF, partly in response to increased competition with the World Bank and other development institutions. Indeed, the Asian Infrastructure Development Bank, led by a political push from China, is sometimes seen as a threat to the IMF's relevance. Along another margin, some studies have suggested that the IMF often competes with private capital markets.

Our concern is that the IMF's increasing involvement in chronic debt crises and in development finance may make it harder to focus on its original mission. While the modern demands on the IMF with nearly 200 members are more diverse than ever before, the old or original needs as defined in the Articles of Agreement remain as compelling as ever. It is true that classic balance-of-payments problems may no longer arise as frequently as they once did because exchange rates are no longer as likely to be fixed or predetermined. However, de jure exchange rate arrangements that promise floating exchange rates are often more flexible on paper than they are in practice, and "fear of floating" has diminished but not disappeared (Levy-Yeyati and Sturzeneger 2007; Ilzeztki, Reinhart, and Rogoff 2016).

There are further reasons why an international lender of last resort remains indispensable. The high levels of international reserves that we observe today should not be taken for granted. International commodity prices can change abruptly and deteriorate over long periods, and so can global financial conditions. Most developing or emerging economies (and not a few advanced ones) have large stocks of debts in foreign currency. Moreover, a domestic lender of last resort faces limitations in emerging markets, particularly those that are highly dollarized (Calvo 2006). In the past, the US Federal Reserve has shown a remarkable willingness to provide dollar liquidity in crisis times, for example when markets froze in 2009. Yet, the Fed's role and mandate are first and foremost domestic. Although there are numerous global and regional development agencies, the world economy lacks a global central bank.

The inherent conflict faced by the IMF is between strengthening its role as an international lender of last resort and the demands of many member countries for serial lending, resulting in repeated programs and a perpetual state of debt "ever-greening." A usual concern in this context is that countries will be tempted to over-borrow if the terms of repayment are so elastic (and after all, both the IMF and the debtors have incentives for ever-greening their loans). However, the point we would like to emphasize instead deals with signaling and stigma.

Many countries appear to welcome (in principle at least) access to liquidity in times of financial stress. The IMF answer is to offer contingent credit lines, which can supplement the self-insurance of countries provided by their holdings of international foreign exchange reserves. However, an international discount window faces many of the same problems faced by discount window facilities of domestic central banks. In a domestic setting, banks often shy away from approaching a central bank's discount window for fear that temporary illiquidity will be mistaken for insolvency by fellow market participants (for discussions of this issue, see Board of Governors of the Federal Reserve 2015; European Central Bank 2015; Bank of International Settlements 2014). As one example, an important object lesson in the Federal Reserve's history comes from the crisis of Continental Illinois National Bank and Trust Company in 1984. In order to meet the demands created by bank runs, Continental Illinois borrowed heavily from the Fed discount window in 1984. The association of the discount window with a failing institution set a precedent of adverse signaling that subsequently led other banks to avoid the discount window for fear they would be deemed as similarly troubled institutions. A policy instrument was damaged and lost.

It seems plausible that countries may worry about stigma in a manner similar to banks. The fact that so much of IMF lending in recent decades is longer term and connected to long-run solvency problems may taint all of its lending. The importance of appearances and signaling, especially how these factors may manifest themselves in times of stress and confusion, should not be underrated. If an IMF program carries a signal of insolvency and is categorically associated with other chronic problems, engaging in a program carries a risk of sending a negative signal about a country's solvency.

If removing development finance and insolvent nations from IMF lending programs is not in the cards, there may at least be merit in more deliberately separating lender-of-last-resort activities from the remainder of what the Fund does. In recent years, the IMF has made efforts to move in this direction. It introduced several new program lines that tilt (albeit broadly) in the direction of behaving like a central bank discount window—that is, being willing to extend a substantial volume of credit on short notice. First, the Rapid Credit Facility (RCF) seeks to provide quick loans with limited conditions to low-income countries facing a balance of payments crisis. Second, the Flexible Credit Line (FCL), enacted in 2009, has the specific goal of reducing the risk of stigma. FCL loans are granted for crisis-prevention and crisis-mitigation in more stable economies—that is, for countries with very strong policy frameworks and before the economy is at severe risk. Third, the Precautionary and Liquidity Line (PLL) is aimed at countries with essentially sound economic fundamentals but with a limited number of vulnerabilities that disqualify them from using the Flexible Credit Line.

At least so far, these discount-window-like programs have not shown much success. The Flexible Credit Line has only had three applicants (Colombia, Mexico, and Poland), while the Precautionary and Liquidity Line has been arranged in two cases only (in the Former Yugoslav Republic of Macedonia and Morocco). Perhaps the limited impact of these programs is due to the fact that issues of stigma remain. This danger of being unable to perform a lender-of-last-resort function because of straying from its original mission is most likely the fuel behind the recommendation of Calomiris and Meltzer (1999) that the IMF act only as lender of last resort and only to countries that meet certain prerequisite standards in banking.

In this journal more than a decade ago, Fischer (1999) touched upon many of the central issues related to the need for an international lender of last resort, particularly in connection with reducing the intensity of financial crises and limiting contagion. We agree with Fischer that a modest element of crisis avoidance may be within reach if there is a more effective lender of last resort. Furthermore, we recognize that the IMF's lending can mitigate the impact of crises on economic development and on long-term growth (on the impact of crises see Cerra and Saxena 2008; Reinhart and Reinhart 2010, 2015).[8]

Concluding Observations: What Lies Ahead?

In the last few years, while advanced economy borrowing from the IMF has reached historic highs, emerging markets have mostly abstained from IMF borrowing.

[8] Using a sample spanning 1870–2014, Reinhart and Reinhart (2015) document that crises are typically associated with lower medium-term growth. Given that the forces for convergence of income across countries are estimated to be slow, an economy that goes off track at the time of a financial crisis may well experience long-lived consequences for its relative economic development—consequences that could have been mitigated by an active and able international lender of last resort.

This has much to do with the favorable external environment that emerging markets faced during much of the past decade: US interest rates were low, declining, and mostly negative after adjusting for inflation; commodity prices were rising markedly; China's investment-led record growth rates fueled the appetite for primary commodity exports; bleak asset returns in advanced economies set off the eternal quest for yields among global investors, favoring emerging markets as an asset class. Good policies helped, too: unlike in prior commodity price booms like the 1970s, many developing country governments managed to avoid heavily pro-cyclical fiscal policies.

But the era of tranquility for emerging markets appears to have ended (at least temporarily). The risks are many. During the last few years, firms and banks in emerging market economies have increasingly succumbed to the temptation of borrowing at low international interest rates while their currencies were either stable or appreciating against the dollar. Current account deficits have reappeared for many of these countries, along with domestic credit booms and currency overvaluation. Moreover, growth began to slow and the US Federal Reserve announced its plans, during spring 2013, to gradually withdraw from its exceptionally accommodative policies. Since then, a sharply appreciating US dollar coupled with significant domestic currency depreciations in many emerging markets have increased external debt burdens.

It is precisely in such an unsettling environment for emerging markets that the IMF may face another wave of demands on its resources. At such a juncture, it appears particularly important that the IMF strengthens its international lender-of-last-resort capabilities and works towards reducing the stigma of IMF lending. The legacy of the recent IMF lending in Europe raises the question of whether the IMF's next round of programs will ratchet up to 10–16 percent of GDP too. If emerging market economies and/or advanced economies in crisis situations were to need and request the same program scale going forward, this would imply more risks for the IMF portfolio and surely increase the likelihood that the IMF ends up lending into insolvency. A good starting point to mitigate these risks would be to strictly apply the debt sustainability criteria prescribed by the IMF. Serial lending to low-income countries and countries with severe debt sustainability problems moves the functioning of the IMF institution quite close to that of development agencies, which is an increasingly crowded field. In our view, the only way to preserve the unique status and the seniority of the IMF is to assure that its lending focuses on the task of providing liquidity quickly to countries in response to short-term financial crisis—that is, acting as a lending source of last resort.

■ *The authors wish to thank Aitor Erce, Gordon Hanson, Enrico Moretti, Vincent Reinhart, Kenneth Rogoff, Timothy Taylor, and Lena Thurau for helpful comments. Maximilian Mandl and Maximilian Rupps provided excellent research assistance.*

References

Bank of International Settlements. 2014. *Re-Thinking the Lender of Last Resort.* BIS Papers no. 79.

Barro, Robert J., and Jong-Wha Lee. 2005. "IMF Programs: Who Is Chosen and What Are the Effects?" *Journal of Monetary Economics* 52(7): 1245–69.

Beers, David T., and Jean-Sébastien Nadeau. 2015. "Database of Sovereign Defaults, 2015." Bank of Canada Technical Report 101, May.

Behrman, Greg. 2007. *The Most Noble Adventure: The Marshall Plan and the Reconstruction of Post-War Europe.* London: Aurum Press Ltd.

Bird, Graham, Mumtaz Hussain, and Joseph P. Joyce. 2004. "Many Happy Returns? Recidivism and the IMF." *Journal of International Money and Finance* 23(2): 231–51.

Board of Governors of the Federal Reserve System. 2015. "The Federal Reserve Discount Window." https://www.frbdiscountwindow.org/.

Bordo, Michael D., and Harold James. 2000. "The International Monetary Fund: Its Present Role in Historical Perspective." NBER Working Paper 7724.

Bulow, Jeremy, and Kenneth Rogoff. 1990. "Cleaning up Third World Debt Without Getting Taken to the Cleaners." *Journal of Economic Perspectives* 4(1): 31–42.

Bulow, Jeremy, and Kenneth Rogoff. 2005. "Grants versus Loans for Development Banks." *American Economic Review* 95(2): 393–97.

Calomiris, Charles W., and Allan H. Meltzer. 1999. "Reforming the IMF." Unpublished paper, Columbia Business School, New York.

Calvo, Guillermo A. 2006. "Monetary Policy Challenges in Emerging Markets: Sudden Stop, Liability Dollarization, and Lender of Last Resort." NBER Working Paper 12788.

Cerra, Valerie, and Sweta Chaman Saxena. 2008. "Growth Dynamics: The Myth of Economic Recovery." *American Economic Review* 98(1): 439–57.

Claessens, Stijn, M. Ayhan Kose, and Marco E. Terrones. 2009. "What Happens during Recessions, Crunches and Busts?" *Economic Policy* 24(60): 653–700.

Cooper, Richard N. 1971. "Currency Devaluation in Developing Countries." Essays in International Finance no. 86, June. Princeton University.

Cruces, Juan J., and Christoph Trebesch. 2013. "Sovereign Defaults: The Price of Haircuts." *American Economic Journal: Macroeconomics* 5(3): 85–117.

Dell'Ariccia, Giovanni, Enrica Detragiache, and Raghuram Rajan. 2008. "The Real Effect of Banking Crises." *Journal of Financial Intermediation* 17(1): 89–112.

Díaz-Alejandro, Carlos F. 1983. "Stories of the 1930s for the 1980s." Chap. 1 in *Financial Policies and the World Capital Market: The Problem of Latin American Countries,* edited by Pedro Aspe Armella, Rudiger Dornbusch, and Maurice Obstfeld. University of Chicago Press.

Díaz-Alejandro, Carlos F. 1984. "Latin American Debt: I Don't Think We Are in Kansas Anymore." *Brookings Papers on Economic Activity*, no. 2, pp. 335–403.

Dreher, Axel, and Nathan M. Jensen. 2007. "Independent Actor or Agent? An Empirical Analysis of the Impact of US Interests on IMF Conditions." *Journal of Law & Economics.* 50(1): 105–124.

Dreher, Axel, Jan-Egbert Sturm, and James Raymond Vreeland. 2009. "Global Horse Trading: IMF Loans for Votes in the United Nations Security Council." *European Economic Review* 53(7): 742–57.

Economist, The. 2008. "The IMF Downsizes: It's Mostly Firing." February 7.

Economist, The. 2009 "The IMF: Battling Stigma: The IMF Is in Search of a Role, and a Happier Reputation." March 26.

Edwards, Sebastian. 1989. "The International Monetary Fund and the Developing Countries: A Critical Evaluation." *Carnegie-Rochester Conference Series on Public Policy* 31: 7–68.

Eichengreen, Barry. 1992. *Golden Fetters: The Gold Standard and the Great Depression 1919–1939.* Oxford University Press.

Erce, Aitor. 2013. "Sovereign Debt Restructurings and the IMF: Implications for Future Official Interventions." Federal Reserve Bank of Dallas, Globalization and Monetary Policy Institute Working Paper 143.

European Central Bank. 2015. "ELA Procedures." Available at: https://www.ecb.europa.eu/mopo/ela/html/index.en.html.

Feldstein, Martin. 1998. "Refocusing the IMF." *Foreign Affairs* 77(2): 20–33.

Fischer, Stanley. 1999. "On the Need for An International Lender of Last Resort." *Journal of Economic Perspectives* 13(4): 85–104.

Gold, Joseph. 1970. *The Stand-By Arrangements of the International Monetary Fund: A Commentary on Their Formal Legal and Financial Aspects.* Washington, DC: International Monetary Fund.

Gourinchas, Pierre-Olivier, and Maurice Obstfeld. 2012. "Stories of the Twentieth Century for the Twenty-First." *American Economic Journal: Macroeconomics* 4(1): 226–65.

Ilzeztki, Ethan, Carmen M. Reinhart, and Kenneth Rogoff. 2016. "Exchange Rate Arrangements Entering the 21st Century: Which Anchor Will Hold?" Unpublished paper. Harvard University.

International Monetary Fund (IMF). 1999. "The IMF's Enhanced Structural Adjustment Facility (ESAF): Is It Working?" September. https://www.imf.org/external/pubs/ft/esaf/exr/.

International Monetary Fund (IMF). 2002. *Evaluation of Prolonged Use of IMF Resources.* Independent Evaluation Office.

International Monetary Fund (IMF). 2013. "Greece: Ex Post Evaluation of Exceptional Access under the 2010 Stand-By Arrangement." Country Report No. 13/156.

International Monetary Fund (IMF). 2014. "The Fund's Lending Framework and Sovereign Debt—Preliminary Considerations." IMF Staff Report, June.

International Monetary Fund. Various dates. *World Economic Outlook.*

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor. 2013. "When Credit Bites Back." *Journal of Money, Credit, and Banking* 45(s2): 3–28.

Joyce, Joseph P. 2005. "Time Past and Time Present: A Duration Analysis of IMF Program Spells." *Review of International Economics* 13(2): 283–97.

Kaminsky, Graciela L., and Carmen M. Reinhart. 1999. "The Twin Crises: The Causes of Banking and Balance-of-Payments Problems." *American Economic Review* 89(3): 473–500.

Killick, Tony. 1995. *IMF Programmes in Developing Countries: Design and Impact.* New York: Routledge.

Kindleberger, Charles P. 1978. *Manias, Panics and Crashes: A History of Financial Crises.* New York: Wiley.

Krueger, Anne O. 1978. *Liberalization Attempts and Consequences.* NBER Books, Cambridge, MA: National Bureau of Economic Research.

Levy-Yeyati, Eduardo, and Federico Sturzenegger. 2007. "Fear of Appreciation." World Bank Policy Research Working Paper 4387.

Mody, Ashoka. 2015. "In Bad Faith." Bruegel Blog Post, July 3.

Mody, Ashoka, and Diego Saravia. 2006. "Catalysing Private Capital Flows: Do IMF Programmes Work as Commitment Devices?" *Economic Journal* 116(513): 843–67.

Moreno, Pablo. 2013. "The Metamorphosis of the IMF (2009–2011)." Bank of Spain Working Paper 78.

Obstfeld, Maurice, and Alan M. Taylor. 2004. *Global Capital Markets: Integration, Crisis, and Growth.* Cambridge University Press.

Reinhart, Carmen M., and Vincent R. Reinhart. 2010. "After the Fall." In Proceedings of the Federal Reserve Bank of Kansas City Economic Policy Symposium "Macroeconomic Challenges: The Decade Ahead," 17–60. Available at: https://www.kansascityfed.org/publications/research/escp/symposiums/escp-2010.

Reinhart, Carmen M., and Vincent R. Reinhart. 2015. "Financial Crises, Development, and Growth: A Long-term Perspective." *World Bank Economic Review* 29(suppl.1): S53–S76.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2009. *This Time Is Different: Eight Centuries of Financial Folly.* Princeton University Press.

Reinhart, Carmen M., and M. Belen Sbrancia. 2015. "The Liquidation of Government Debt." *Economic Policy* 30(82): 291–333.

Reinhart, Carmen M., and Christoph Trebesch. Forthcoming. "Sovereign Debt Relief and its Aftermath." *Journal of the European Economic Association.*

Rose, Andrew K. 2005. "One Reason Countries Pay Their Debts: Renegotiation and International Trade." *Journal of Development Economics* 77(1): 189–206.

Sachs, Jeffrey. 1998. "The IMF and the Asian Flu." *The American Prospect* no. 37 (March–April).

Schadler, Susan. 2013. "Unsustainable Debt and the Political Economy of Lending: Constraining the IMF's Role in Sovereign Debt Crises." CIGI Paper no 19.

Schadler, Susan. 2014. "The IMF's Preferred Creditor Status: Does It Still Make Sense after the Euro Crisis?" CIGI Policy Brief 37, March.

Schlegl, Mattias, Christoph Trebesch, and Mark. L. J. Wright. 2015. "The Seniority Structure of Sovereign Debt." Unpublished paper.

Stone, Randall W. 2004. "The Political Economy of IMF Lending in Africa." *American Political Science Review* 98(4): 577–91.

Subramanian, Arvind. 2015. "How The IMF Failed Greece." *Project Syndicate*, August 13.

Thacker, Strom C. 1999. "The High Politics of IMF Lending." *World Politics* 52(1): 38–75.

World Bank. No year. *International Debt Statistics.* A database. Washington DC: World Bank.

Journal of Economic Perspectives—Volume 30, Number 1—Winter 2015—Pages 29–52

The IMF's Unmet Challenges

Barry Eichengreen and Ngaire Woods

The International Monetary Fund is a controversial institution whose interventions regularly provoke passionate reactions. On one side are those like Krugman (1998) and Nissan (2015) who argue that the IMF is an "indispensable institution." On the other are critics who object that the Fund is unrepresentative, inefficient, and an engine of moral hazard and conclude that the world would be better off without it (International Financial Institutions Advisory Commission 2000; Global Exchange 2015).

Dispassionate analysis should start with articulation of the problems that the IMF is designed to solve. It should then turn to the question of whether the Fund is appropriately structured to solve those problems—and whether it can do so without creating other equally serious problems through its own actions.

We will argue that there is an important role for the IMF in helping to solve information, commitment, and coordination problems with significant implications for the stability of national economies and the international monetary and financial system. In its role as a trusted advisor to governments, the Fund can apply lessons from the experience of other countries, basing its analysis on information that national authorities are not inclined to share with other interested parties such as rating agencies and investment banks. It can raise awareness of cross-border spillovers of

■ *Barry Eichengreen is the George C. Pardee and Helen N. Pardee Professor of Economics and Political Science, University of California, Berkeley, California. Ngaire Woods is Dean of the Blavatnik School of Government and Professor of Global Economic Governance, University of Oxford, Oxford, United Kingdom. Their email addresses are eichengr@econ.berkeley.edu and ngaire.woods@bsg.ox.ac.uk.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.1.29

doi=10.1257/jep.30.1.29

policies that governments would otherwise have little incentive to acknowledge and encourage mutually advantageous policy adjustments to internalize those externalities. As emergency lender, it can prevent cash-strapped governments from having to resort to policies that could endanger domestic and international financial stability. By combining lending with advice, it can strengthen the hand of national authorities seeking to put in place stability-enhancing reforms.

In executing these functions, the effectiveness of the IMF, like that of a football referee, depends on whether the players see it as competent and impartial. We will argue that the Fund's perceived competence and impartiality, and hence its effectiveness, are limited by its failure to meet four challenges. Certain of these challenges are conceptual: the trusted advisor doesn't always know what to advise. Others are practical: as a result of its organization, the IMF's impartiality is called into question. All four of the challenges threaten the legitimacy of the institution and therefore its capacity to execute its core functions.

The first unmet challenge is how to organize the *surveillance* through which the IMF "monitors the economic and financial policies of its 188 member countries. . . highlights possible risks to stability and advises on needed policy adjustments" (IMF 2015a). The view of its founders was that the Fund should focus like a laser on the prerequisites for stable exchange rates and engage in ruthless truth-telling when those prerequisites were not met.[1]

But over time, surveillance moved away from the exchange rate issues that were central to the Fund's original mandate. Surveillance expanded to where it now encompasses virtually anything and everything with implications for economic and financial stability. Blunt truth-telling, meanwhile, remains more the exception than the rule. It may be unrealistic to expect that the Fund should have anticipated and warned of the US subprime crisis, the global financial crisis, and the Greek debt crisis. But the IMF batted 0 for 3 on these three events, which suggests that its capacity to "highlight risks to stability" leaves something to be desired.

Second, there is confusion about what kind of *conditionality* should be attached to IMF loans. Conditionality refers to policy commitments by governments made in return for receiving assistance. But there is disagreement about how many and what kind of commitments to require. Why should the IMF impose conditions on a government that the latter does not view as in its self-interest? How can it do so without infringing on the member's sovereignty, which the IMF as a multilateral institution is obliged to respect? Since governments are sovereign, why should one expect them to meet conditions designed to achieve goals and objectives that they do not fully embrace? Conversely, if the Fund and a government have the same objectives, why are conditions needed in the first place?

[1] The phrase "ruthless truth-telling" is commonly attributed to John Maynard Keynes in connection with negotiations to establish the institution. In fact, the phrase is due to Keynes, but in a different (earlier) context, in a letter written to Jan Smuts concerning the post–World War I peace negotiations at Versailles. See Johnson (1977, p. 8).

Third, there is disagreement about the IMF's role in *the management of sovereign debt crises*. Multiple stakeholders, significant transactions costs, and the absence of an internationally agreed legal framework for resolving sovereign debt crises create coordination problems that constitute a prima facie case for the involvement of a multilateral institution like the IMF. But there is confusion about the form that involvement should take. Officials give lip service to the idea that the IMF should provide emergency liquidity assistance when a government's debt is sustainable but private investors are unable to coordinate the provision of such liquidity, and only facilitate a debt restructuring when the debt burden is not sustainable. However, such statements beg deep questions about whether the concept of debt sustainability is meaningful and, if it is, whether the Fund is capable of determining the sustainability of a sovereign's debt in practice. Reflecting these uncertainties, the IMF has regularly continued lending for too long and put off the restructuring decision. This IMF pattern allows private investors to cut their losses and creates moral hazard. When the restructuring finally comes, it is more expensive for the country and more disruptive to the economy.

Fourth, *governance problems* raise questions about the Fund's impartiality. Some members have disproportionate voice, enabling them to sway decision making in directions consistent with their national interest even when those directions are at odds with both the interest of IMF membership and the stability of the international monetary and financial system broadly defined. Other members are inadequately represented and consequently see the Fund's decisions and advice as neglecting their interests. Both groups are therefore reluctant to give IMF staff and management more autonomy in designing programs and choosing tactics.

The IMF's failure to more successfully meet these four challenges causes its critics to question its competence and impartiality and—returning to the earlier analogy—to see it as a discredited referee. In the language of political theory, the Fund's failure to meet these challenges leads members to question its *legitimacy*. Political scientists define legitimacy as the presumption "by an actor that a rule or institution ought to be obeyed" (Hurd 1999; see also Beetham 1991, Dahl 1994, Scharpf 1997, and Flathman 2012). The legitimacy of a multilateral institution such as the IMF thus determines the willingness of agents—governments and their constituencies in the present context—to accept its recommendations and bend to its authority.

Political theorists argue that legitimacy has two sources. "Output legitimacy" refers to public assessments of the quality of the institution's performance. If its results are good, then the institution's advice is respected and agents are willing to acknowledge its authority. "Input legitimacy" refers to the process through which decisions are reached and power is exercised. If that process gives voice to and empowers relevant stakeholders appropriately, and if the interests of different parties are properly weighed—if the decision-making process is even-handed and subject to checks and balances—then the institution's advice is again more likely to be accepted.

The first three challenges—concerning the quality of surveillance, the relevance of conditionality, and the utility of the Fund's approach to debt problems—create

doubts about the IMF's output legitimacy. The fourth challenge, the Fund's failure to adopt a system of governance that gives appropriate voice to different stakeholders, raises questions about its input legitimacy. These problems of legitimacy will have to be addressed in order for the IMF to play a more effective role in the 21st century.

Surveillance

Article IV of the 1944 Articles of Agreement, the IMF's original constitution, obliged members to maintain fixed par values for the foreign exchange rate of currencies and to obtain the Fund's concurrence before changing them. Compliance with this obligation was the touchstone of the institution's efforts to monitor and advise on national policies.

With the collapse in 1971–73 of the Bretton Woods System of pegged but adjustable exchange rates, IMF surveillance (like those exchange rates themselves) was cast adrift. Discussions about how to proceed culminated in a 1977 decision on surveillance that shifted the goal posts. Where the original Article IV had alluded to a system of stable exchange rates, the 1977 decision referred instead to a stable system of exchange rates. This minor change had major consequences: it reoriented surveillance from the stability of exchange rates to the stability of the "system"—and, by implication, to the policies influencing that stability.

Reviews of national economies were therefore expanded to consider whether monetary, fiscal, and financial policies were consistent with "stability" broadly defined. IMF staff met with government officials annually in so-called Article IV consultations. They gathered information on the national economy and provided analysis of the external environment facing the country. They assessed the mutual consistency of their forecasts for individual countries and attempted to gauge threats to the stability of the international monetary and financial system in a growing list of exercises and reports: a *World Economic Outlook*; an *International Capital Markets Report*; a *Global Financial Stability Report;* a *Fiscal Monitor*, a set of *Financial Sector Assessments* integrating macroeconomic and financial surveillance; and a series of *Spillover Reports* concerned with how the policies of large countries affect the rest of the world.

Risks and spillovers have thus become the central focus of IMF surveillance (IMF 2014a), which is consonant with the rationale for the existence of the Fund as we frame it above. Risks to economic and financial stability are complex. They are difficult for national authorities to assess. They will depend on conditions not just at home but also in other countries, information which is difficult to assemble. There is thus a role for the Fund in assembling and conveying information about global economic and financial conditions and imparting lessons learned from its experience in other countries to members in need of this expertise. Those members, for their part, are more likely to share sensitive information with a multilateral institution like the Fund than with interested parties like rating agencies and investment banks.

In addition, elected officials with a limited tenure in office often have short horizons and consequently high discount rates, which blunts their incentive to invest

in analyzing and heading off future risks. They are likely to take inadequate account of the cross-border spillovers of their policies insofar as these do not directly impact domestic constituencies in the near term.

On all these grounds, there is a role for IMF surveillance as corrective. The underlying rationales are not unlike those motivating the stress tests used by national supervisors to gauge risks to the stability of domestic banks, with the subject in this case being the national balance sheet rather than a bank balance sheet. That the Fund undertakes both bilateral and multilateral surveillance (surveillance of individual countries in the first instance and of groups of countries and the global system in the second) is then analogous to the pursuit of both micro-prudential and macro-prudential policies by domestic authorities responsible for financial stability (Tucker 2015).

Viewed in this light, the Fund's failure to sound louder warnings in the run-up to the subprime crisis in the United States, the banking crises in Iceland, Ireland, and other countries, and the sovereign debt crisis in Greece is troubling. The IMF had never conducted a Financial Sector Assessment Program review of the United States before the crisis. (The first such review of the US economy took place in 2009–2010.) In a financial sector assessment of Ireland released in August 2006, an IMF team concluded that "the financial system seems well placed to absorb the impact of a downturn in either house prices or growth more generally." While an IMF staff mission to Greece, in its concluding statement in May 2009, acknowledged the likelihood of a slowdown in growth and the desirability of fiscal consolidation and structural reform, its assessment of the state of the Greek banking system was sanguine ("[t]he banking system appears to have enough buffers to weather the expected slowdown"). It projected a budget deficit for 2009 of "at least 6 per cent of GDP," less than half the level acknowledged by the new Greek government at the end of the year.

In a seminal analysis of IMF surveillance, Mussa (1997) pointed to several explanations for such failures. Some of the issues stem from the practical limitations of economic analysis. "The economics profession has not discovered the magic formula that assures rapid and steady economic growth, low inflation, financial stability, and social progress," as Mussa disarmingly put it. There is the analytic challenge of integrating macroeconomic and financial analysis. There is the organizational challenge of overcoming informational silos within the institution. There is the struggle to staff up an agency traditionally dominated by macroeconomists with individuals possessing the specialized skills of bank regulators and schooled in the working of financial markets. There are the special difficulties of conducting surveillance reviews of countries that are members of a monetary union and therefore lack some of the instruments of national economic policy. There is the worry that warning of economic and financial weaknesses could precipitate a loss of confidence and a crisis that would not otherwise occur. There is the challenge of deciphering the state of the economy though the practice of flying visits to a country several times a year.[2]

[2] More intensive engagement, on the other hand, creates the danger of IMF staff going native—of identifying with their interlocutors and believing what they are told—and the resulting tendency of growth forecasts to err in the direction of overoptimism (de Resende 2014).

Table 1
Deletions in Article IV and UFR (Use of Fund Resources) Staff Reports (2012–2014)

	2012		2013		2014	
Reports by group	*Number published*	*% of reports with deletions*	*Number published*	*% of reports with deletions*	*Number published*	*% of reports with deletions*
All Article IV and Use of Fund Resources reports	178	18%	179	17%	180	19%
Advanced markets	**35**	**23%**	**41**	**32%**	**39**	**23%**
European Union	23	26%	31	39%	26	12%
Other Europe	4	25%	4	25%	3	0%
Rest of advanced markets	8	13%	6	0%	10	60%
Emerging markets	**73**	**26%**	**73**	**19%**	**80**	**26%**
Developing countries	**70**	**7%**	**65**	**5%**	**61**	**8%**

Source: IMF (2015b, p. 6).
Notes: Governments can request that passages of IMF reviews be deleted from the final report. The numbers in the table refers to documents considered by the Board during the calendar year, and published within six months of the end of the calendar year: for example, the publication rate for 2014 refers to the documents discussed by the Board in 2014 and published by June 30, 2015. The data are based on definitions in World Economic Outlook reports, including the new definition of Low-Income Developing Countries established in 2014 (IMF 2014c; http://www.imf.org/external/np/pp/eng/2014/060314.pdf), and on discussions held with members' territories and currency unions in the context of Article IV consultations with their constituent members.

A concern expressed in surveys of the Fund's emerging market members is that advanced countries with disproportionate voice and power in the IMF can bias surveillance reviews in their favor or avoid those reviews entirely. Some surveillance activities are voluntary, like Financial Sector Assessment reviews. Members must first request that they take place, and powerful countries like the United States can resist pressure to volunteer. Governments can request that passages of IMF reviews be deleted from the final report, based on statements that information is market-sensitive or that it does not refer to a policy intervention. These deletion rates are generally higher for advanced and emerging economies, as shown in Table 1.

In their most recent surveillance review, IMF (2014a) staff acknowledged that "key messages, including on spillovers, are sometimes buried deep in reports. External studies and interviews suggest that this may reflect additional internal pressure and scrutiny associated with surveillance of systemic economies."[3] In a survey by the Fund's Independent Evaluation Office (2013, p. 26), nearly 60 percent of

[3] Systemic economics are to the global economy what "global systemically important financial institutions," or G-SIFIs, are to the global financial system. The Financial Stability Board published a list of G-SIFIs in 2011, and the recent revision of the Basel Accord (Basel III) is intended to target the risks they pose for global financial stability.

mission chief respondents who worked on advanced countries acknowledged "pressure to dilute the candor of staff reports in order to avoid upsetting the country authorities" (Callaghan 2014, p.15). Blunt truth-telling about risks and spillovers evidently remains easier in theory than in practice.

To some of these criticisms, the appropriate response is: Don't ask too much. As Mussa (1997) observed, the IMF shouldn't be expected to predict all crises. To other criticisms, the appropriate response is: The Fund must try harder. It should focus its resources more effectively where the case for surveillance is strongest, namely, where risks to stability are *serious*, where information and commitment problems are *acute*, where spillovers are *pronounced*. That said, evidence on deletion rates, questions about evenhandedness, and the difficulty of blunt truth-telling suggest that reforms of IMF procedure and governance would help to strengthen the surveillance function.

Exchange Rates and Capital Flows as a Case in Point

The IMF was founded to address the problems created by disorderly and unstable exchange rates in the 1930s and to discourage exchange rate policies that might injure national economies and disrupt the operation of the international system. Article I of the Articles of Agreement describes the central purpose of the Fund as "to promote exchange stability, to maintain orderly exchange arrangements among members, and to avoid competitive exchange depreciation." The 1977 shift in the meaning of surveillance, noted earlier, acknowledged the demise of the Bretton Woods System of pegged but adjustable exchange rates but continued to place exchange rates at the center of surveillance.

The question is whether exchange rates still deserve that singular position in surveillance today and, if so, what form monitoring and advice should take. Lack of clarity on the answers, we believe, is a significant part of what creates doubts about IMF surveillance in practice. This is an area where, for historic reasons, the Fund is supposed to possess special expertise. Lack of clarity on the rationale and terms of exchange rate surveillance is therefore especially damaging to perceptions of its competence.

An emphasis on risks and spillovers suggests focusing on the exchange rate in two specific instances: 1) in small open economies where the exchange rate has an especially powerful effect on inflation, growth, and financial-market conditions—where, as the point is sometimes put, the exchange rate is the single most important price in the economy (Eichengreen 1994, p. 2); and 2) in large ("systemic") economies whose exchange rate policies are a source of significant cross-border spillovers—where it really matters that the exchange rate is the *relative* price of two currencies. In countries that fall between these stools, in contrast, focusing on the exchange rate does more to diminish the effectiveness of surveillance than enhance it.

However, what it means to have "stable" exchange rates in a modern global economy is not altogether clear. At times, the Fund has encouraged members to

move toward more flexible, market-determined rates, albeit with less than full success. At other times, it has encouraged countries with rigid exchange rates to maintain them, lending to governments when their currencies were under pressure, as in Argentina in 2000–2001 (an episode that did not end happily). It has urged large countries like the United States and China whose exchange rates affect the stability of the international financial system to allow their currencies to move to levels "consistent" with the stability of that system without providing much specificity on what is meant by consistent. It has failed to adapt its surveillance and lending adequately to the circumstances of monetary unions whose members lack a national currency.

In response to such complaints, the IMF commissioned a survey and review by its Independent Evaluation Office (2007), which concluded that the Fund was "simply not as effective as it needs to be in both its analysis and advice, and in its dialogue with member countries." Internal discussion then produced another "Decision on Bilateral Surveillance" in 2007. This focused on external stability as the organizing principle for bilateral surveillance, defining external stability as "a balance of payments position that does not, and is not likely to, give rise to disruptive exchange rate movements." The document concluded that members should "avoid exchange rate policies that *result* in external instability, regardless of their purpose" and that IMF monitoring and advice should be geared to this end.

But this definition of stability—avoid steps that cause instability—is not very useful as a basis for concrete advice. In attempting to operationalize it, the Fund seems to have embraced the "bipolar view" of exchange rate regimes (Fischer 2001). In this view, countries with financial markets that are relatively open and are therefore exposed to large international capital flows should adopt either flexible market-determined rates, a la Chile, or rigid pegs in the manner of Latvia (before it joined the euro area in 2014).[4] Only countries with significant restrictions on international financial transactions and therefore insulated from large capital flows should contemplate intermediate options. Operationally, of course, this still begs the question of how open is "open" and how flexible is "flexible." One has the feeling that IMF economists, when giving advice on exchange rate policy, are still flying by the seat of their pants.

Insofar as countries have been moving in the direction of financial opening, one would expect to see the IMF recommending flexible exchange rates with more frequency. A review of documents by the Independent Evaluation Office (2006, p. 7) confirmed that the Fund's advice on exchange rate policy had indeed shifted in favor of greater exchange rate flexibility. Takagi et al. (2014) found that about half of all programs between 2008 and 2011 called for greater exchange rate flexibility.

[4] In its 2008 program with Latvia, negotiated when the currency and the country were under duress, the IMF agreed to support the government's continued maintenance of a rigid peg to the euro. On the other hand, Chile has intervened in the foreign exchange market only on rare occasions (Claro and Soto 2013).

The alternative for countries with increasingly open capital accounts is to hold their exchange rates fixed using a rigid currency peg, a currency board, or membership in a monetary union. Most recent movement in this direction has taken the form of additional members of the European Union adopting the euro. One would therefore expect to see the IMF offering pointed advice on whether euro adoption is appropriate for a country's circumstances and on how to adjust domestic and euro-area-wide policies so as to make that arrangement operate smoothly.

Given that euro experience has been fraught with difficulty, it is striking that the IMF did not sound louder warnings. It did not warn of the problems of monetary union without banking and fiscal union, despite the fact that seminal work on these points was done by IMF economists. In hindsight, it is clear that the institution was too disposed to accede to the preferences of its European members, who account for a substantial share of voting power and financial contributions to the Fund.

Once the euro was established, IMF surveillance then took inadequate account of large intra-area capital flows that laid the basis for the crisis that erupted in late 2009. It failed to highlight the exposure of euro area banks to sovereign risk and to flag how such risks were heightened by the operation of the monetary union. Despite establishing a review aimed at the euro area as a whole, national Article IV consultations remained the main focus of surveillance, preventing the Fund from obtaining and enunciating a clear picture of threats to the stability of the monetary union (Pisani-Ferry, Sapir, and Wolff 2011, p. 15).

The IMF then participated in a rescue program for Greece, although it contributed only one-third of the finance and therefore had limited influence over its design. It bowed to the wishes of the European Commission and the European Central Bank, its partners in the program, to avoid an early debt restructuring. Its intervention did not restore market confidence, and doubts about Greece's financial prospects spilled over to other euro-area economies. Greece itself experienced a severe recession, and its debt to private creditors was finally restructured in 2012. As a participant in this decision-making Troika, the IMF hardly covered itself in glory.

Exchange rate and capital-flow-management policies are closely bound together, as already noted. In the 1990s the IMF consistently warned against the use of capital controls except in the most exceptional circumstances. The advice at the time was that controls were at best ineffective in managing the stability risks posed by international capital flows (Schadler, Carkovic, Bennett, and Kahn 1993); at worst, they discouraged necessary macroeconomic adjustments. Controls also ran counter to the presumption that financial development and international financial liberalization went hand-in-hand and that financial development was important for economic growth. These views were most forcefully advanced in the second half of the 1990s when, reflecting pressure from the US government, IMF officials mooted the possibility that capital account convertibility—that is, the removal of all significant restrictions on the ability of a firm or individual to exchange currencies whenever they wish—might be made a statutory obligation of members.

The Asian financial crisis of 1997–98 raised doubts about this approach. Volatile and poorly regulated capital flows—inflows in the run-up to the crisis and then

outflows once it was underway—played a prominent role. In the wake of the crisis, a number of academic and policy studies, including some by the IMF itself, cast doubt on the presumption that capital account liberalization contributed more to financial development and growth than to stability risks (Kose, Prasad, Rogoff, and Wei 2006; Prasad and Rajan 2008). Policymakers in emerging markets, their views colored by first-hand experience with those risks, had long been skeptical of full capital account liberalization, and they now reasserted their position. Those officials gained a mechanism for sharing and expressing their views with the creation of new venues outside the Fund such as the Group of Twenty representing the world's largest national economies, which issued sympathetic studies of the role of capital controls as macroprudential policy tools (Group of Twenty 2011). The IMF then moved toward their position, partly, it has been argued, in order to avoid losing influence and jurisdiction over such matters (Gallagher 2015).

The global financial crisis came next. After-the-event analysis, including the IMF's own, showed that countries with capital controls in place before the crisis fared better (Ostry, Ghosh, Habermeier, Chamon, Qureshi, and Reinhardt 2010). Officials from emerging markets who were skeptical of the merits of capital account liberalization became even more assertive once the crisis cast a pall over the policy advice of the advanced economies (Grabel 2015). Iceland, itself an advanced country, experienced large capital inflows and outflows and resorted to controls as a crisis management expedient. This served as a reminder that the problems created by capital flows and recourse to controls were not limited to developing countries.

The result was a series of staff studies constituting a "new institutional view" of capital controls, endorsed by the Executive Board (IMF 2012). (The Executive Board is the IMF's day-to-day governing council; more on this below.) This new view acknowledged that capital flows have substantial benefits but also risks, and that capital-flow-management measures, including controls, "can be useful," although such measures should not substitute for warranted macroeconomic adjustments. The document pointed to circumstances justifying controls, taxes, and financial regulations affecting capital inflows prior to crises (such as when capital inflows heighten risks to domestic and international financial stability that cannot be adequately neutralized by conventional macroeconomic and prudential policies). Equally, it pointed to circumstances justifying their use in response to outflows and crises (to instances when controls might be used to avert collapse of an exchange rate, the depletion of reserves, and restrictions on the operation of the banking and financial system). Gallagher and Tian (2014), analyzing Article IV reports, conclude that the change in staff guidance resulted in a significant change in IMF assessments and advice regarding capital controls.

The crunch will come when the IMF's new view on capital controls brings the institution into conflict with its largest shareholders. Specifically, it is not clear how the Fund will proceed when advising governments with whom the United States is attempting to conclude trade and investment treaties that would limit a country's recourse to capital controls and that lack temporary safeguard provisions "compatible with the IMF's own approach" (IMF 2012).

Lending and Conditionality

In the Bretton Woods era, IMF loans were extended to support the maintenance and orderly adjustment of exchange rate pegs. The Fund lent when a government, lacking adequate foreign currency reserves, needed temporary financial assistance to defend the existing or newly established value of the exchange rate. The conditions attached to those loans consequently focused on monetary and fiscal actions needed to render policies compatible with the declared exchange rate parity.

With the shift in focus from the maintenance of stable exchange rates to the maintenance of stable economic and financial conditions broadly defined, IMF lending programs—Stand-By Arrangements (SBAs) as they are technically known—were extended not merely to help a government defend the particular level of the exchange rate, but to enable it to take various steps needed to maintain economic and financial stability and, equally, to avoid having to take other steps, like suspending the operation of banks and financial markets, in response to a crisis. As the IMF (2015c) puts it in its Stand-By Arrangement Factsheet, the Fund provides Stand-By Arrangements "to respond quickly to countries' external financing needs, and to support policies designed to help them emerge from crisis and restore sustainable growth."

As the range of considerations in response to which the Fund made loans expanded, so too did the conditions it attached to its loans. These conditions came to encompass a panoply of structural, institutional, and procedural measures ranging from public enterprise privatization and pension reform to labor market liberalization and changes in tax administration. This approach reached its apogee in the Asian crisis of 1997–98. It was epitomized by the program for Indonesia, which featured a veritable Christmas tree of conditions (including dismantling the plywood export cartel, eliminating the government monopoly on trade distribution, and removing state subsidies for motor vehicle production). The program was controversial, and the government fell—two events that were seen as not unrelated.

The result was a backlash against expansive conditionality as intrusive, ineffectual, and counterproductive and a 2002 decision by the Executive Board encouraging a more "focused" approach. But that decision, in turn, raised additional questions. Conditions focused on what? Should structural conditions always be avoided, or can structural distortions be the principal source of economic imbalances and even crises, in which case only structural conditions can address the root causes of instability?

Answering these questions requires stepping back and asking why IMF conditionality is needed in the first place. Conditions are attached to loan contracts when there is a conflict of interest between the lender and borrower—when the lender wants the borrower to focus on actions that maximize the likelihood of repayment while the borrower is inclined to give weight to other goals. Thus, one role of IMF conditionality is to "safeguard members' assets" (as maximizing the likelihood of repayment is described in Fund-speak).

But defaults on IMF loans are rare, since defaulting threatens to antagonize important trading partners and to disrupt access to a valued source of emergency

credit. To the extent that this rationale applies, moreover, it suggests prioritizing actions by the borrower that can be completed within 12 to 24 months, the duration of a typical IMF stand-by arrangement, where in practice many elements of IMF conditionality, such as "combatting corruption and strengthening rule of law" (as in Ukraine's recent Stand-By Arrangement) can be implemented and produce results only over much longer periods.[5]

In some sense, the very notion of a conflict between the IMF and a government is problematic. As a condition for lending, the IMF insists that governments should embrace its goals—that governments should take "ownership" of their programs. Drazen (2002) defines ownership as "the extent to which a country is interested in pursuing reforms independent of any incentives provided by multilateral lenders." But, as he proceeds (p. 41) to note: "Conditionality makes little or no sense if there is full ownership, but it also makes no sense if there is no ownership." The IMF, in other words, has no reason to impose requirements on a government if that government is fully committed to meeting those requirements. Equally, however, there is no purpose if the government of a sovereign state is opposed to those requirements, since the implication of sovereignty is that those conditions have no prospect of being met.[6]

Perhaps there is uncertainty about the extent of ownership, in which case conditions can be useful as a signal. Investors may be uncertain about whether a government is committed to reform. Requiring conditionality in return for a loan may enable investors to distinguish which governments are truly committed. This will give the Fund's shareholders, who fund the loan, confidence that they will be paid back. In addition, an IMF loan conditioned on key reforms, by sending this signal, may "catalyze" private capital inflows or stem capital flight (Ghosh, Lane, Schultze-Gattas, Bulír, Hauman, and Mourmouras 2002).[7]

Or perhaps conditions are useful when the IMF and a government share priorities but a domestic constituency disagrees. Lending, coupled with policy conditions, may tip the balance between outcomes when its extension directly affects the welfare of the dissenting group or strengthens the bargaining power

[5] In fact the repayment period can be longer (three to five years) than the duration of the Stand-By Arrangement itself. In addition, the IMF has lesser-used windows like the Extended Fund Facility through which it lends for longer periods to countries with chronic problems. But the tension flagged in the text remains.

[6] The IMF has in fact established two lightly conditioned facilities, a Flexible Credit Line (FCL) for "very strong performers" where safeguarding resources is a nonissue, and a Precautionary and Liquidity Line (PLL) for "countries with sound policies" where it is almost a nonissue. The FCL can be drawn without conditions, the PLL subject to only very limited conditionality. In the language of the text, countries that qualify to borrow from these facilities can be thought of as "fully" and "almost fully" committed to necessary adjustment policies. But there has been little inclination to prequalify for these facilities (only five countries have applied despite the IMF's repeated efforts to make the facilities attractive), and exactly zero appetite to borrow from them, presumably for reasons of stigma. Evidently, countries that don't need to borrow from the Fund can do so without conditions, while countries that need to borrow are unable to do so without conditions.

[7] For this signaling story to be consistent, a variety of ancillary conditions have to be met, as in the famous education-as-signaling model of Spence (1973).

of the authorities. The government and a majority of domestic stakeholders may agree, for example, that cuts in public sector salaries are needed for debt sustainability and the restoration of financial market access, while public employee unions strenuously resist. A loan from the IMF conditioned on public sector pay cuts may then strengthen support for the government's policies among the public at large and moderate the opposition of public sector unions by allowing the adjustment to be phased in more gradually.

Building coalitions for reform or favoring general over special interests is delicate business politically. The more delicate the business, the more important it is that the IMF be seen as evenhanded. This suggests that expansive conditionality, where the areas subject to reform are not tightly linked to the IMF's mandate, is problematic insofar as it creates scope for powerful shareholders to enlist the institution in advancing their national interests. The US Executive Director at the IMF is obliged by Congress to support a range of structural policies in other countries, from promoting economic deregulation to privatizing industry. The US Treasury lists its successes in pressing for such policies in its reports to Congress: US Department of Treasury (2000, pp. 12–13), for example, reports success in pressing for accelerated privatization in Indonesia, Romania, Bulgaria, Nigeria, and the Gambia. This leads governments that are the subject of Fund programs to question the motivations for such policy advice and challenge its legitimacy. It also suggests that governance reform that addresses doubts about whether conditionality is attuned to the welfare of the crisis country and not just the interests of the IMF's principal shareholders would enable the Fund to go about its business more efficiently.

What conditions exactly are appropriate in light of these rationales? The simple answer is that conditionality should focus on policy reforms central to the borrower's ability to restore financial stability and economic growth and exit its IMF program on time. While in some cases these will be macroeconomic policies, in others, they will be financial and structural policies. From this perspective, calls for the IMF to return to earlier practice where conditionality focused exclusively on macroeconomic policies, and was therefore in some sense "less intrusive," are fundamentally misplaced.

Within these categories, should the IMF focus on a short list of key policy adjustments or long menus of macroeconomic, financial, and structural conditions? Here there is an analogy with the "big bang" and" gradualist" approaches to stabilization and adjustment in the transition economies of Central and Eastern Europe in the 1990s. (The literature on this debate is large; useful starting points include Dewatripont and Roland 1992, Roland and Verdier 1994, and Wei 1997.) Proponents of the big bang approach argued that policy reforms are complementary—that the gains from a given reform will be greater in the presence of others. They suggested that it was easier to build a reform-minded coalition when the adjustment burden was shared widely. Gradualists, in contrast, argued that undertaking a wide range of adjustments simultaneously could be disruptive and create unnecessary costs. Attempting to implement multiple reforms could overload the political system; it might foster the development of a broad-based coalition opposing reform. In an environment of uncertainty,

targeting a subset of key reforms is more feasible politically. Society will then be better able to learn about the costs and benefits of further reform. This list of countervailing arguments suggests that there is no general answer to the question of whether lists of IMF program conditions should be long or short.

IMF conditionality will always be criticized. Compliance is painful, especially when costs are front-loaded and benefits are deferred. Powerful special interests will inevitably be opposed. Even focusing on key reforms central to the borrower's ability to restore economic growth and, at the same time, to exit its Fund program—the two obviously legitimate goals of conditionality—is easier said than done. Strengthening the police, judiciary, and tax administration, for example, will be priorities for an IMF that believes economic growth depends on rule of law, but laying off judges, policemen, and tax administrators may be necessary to balance the budget and ensure that the Fund is repaid at the end of three years.

Overall, the IMF needs to specify exactly what its policy conditions are designed to achieve and how exactly they are designed to achieve it. Doing so disciplines the design of those conditions; it encourages a more focused conditionality that avoids overloading the political system. And governance reform that addresses complaints about whose interest is behind the conditions will enhance the perceived legitimacy of the Fund's interventions and the prospects for country ownership and compliance.

Resolution of Sovereign Debt Problems

The IMF has been embroiled in sovereign debt problems continuously in recent years. Procedures for restructuring sovereign debts are ad hoc and imperfect, creating uncertainties for investors and governments.[8] The number and diversity of stakeholders and the absence of an internationally agreed legal framework virtually cry out for a multilateral entity to coordinate negotiations (IMF 2003). The Fund is a logical such entity, since unsustainable debts can threaten domestic and international financial stability and, consequently, are central to its mandate.

But deciding when and how sovereign debts should be restructured poses challenges. Most fundamentally, there is the challenge of determining whether a country's debt is sustainable—whether, given reasonable adjustment effort and short-term liquidity assistance, the government is able to pay. Assessments of debt sustainability require assumptions about future interest rates and growth rates, which are intrinsically uncertain over the long horizons relevant to sovereign debt contracts. Ability to pay is not easily distinguished from willingness to pay. In other

[8] It is sometimes argued that procedures for restructuring *should* be ad hoc and imperfect, since if they were perfectly predictable it would become too tempting for debtors to restructure their obligations. But the point, developed below, is that a strong argument can be made that they are too imperfect, causing restructurings to be unnecessarily delayed. An earlier review of these issues in this journal is Eichengreen (2003).

words, a government's decision on whether to make or suspend payments depends on political as well as economic considerations.

How should the IMF factor in these political considerations when deciding whether to provide an emergency loan or instead insist on a restructuring? Again there are unlikely to be hard and fast answers. IMF officials have no choice but to use their best judgment, which in turn makes it important that their judgment is seen as seeking to enhance the stability of the member and the international system, and that it does not simply reflect the narrow interest of a group of powerful stakeholders. The Fund's decision not to insist on a Greek debt restructuring in 2010 is a widely cited example of what not to do. In that instance, the French and German governments were concerned that restructuring Greece's debt might adversely affect their banks, and they were powerful enough to get their way (Blustein 2015; Orphanides 2015). Again, this observation points to governance as a factor affecting the legitimacy of IMF decision making.

While Greece is a prominent case in point, the criticism that the IMF has often waited too long to recommend sovereign debt restructuring is more general (IMF 2013b provides additional examples). In part, this criticism is informed by 20/20 hindsight when matters have turned out poorly: evaluating debt sustainability, as already noted, is not a matter of black or white. In addition, there can be uncertainty about whether the restructuring will be executed smoothly or the offer to exchange old bonds for new ones of lesser value or longer maturity will be rejected by investors, causing the country to lose financial market access and incur other costs. There are also worries about spillovers—concerns over how a debt restructuring might affect the market access and financial systems of other countries.[9]

For all these reasons, there is a temptation to provide emergency finance, some of which goes to keeping debt service current, while hoping that good news turns up. When it doesn't, and restructuring is unavoidable—Table 2 lists countries that experienced restructuring following an IMF program—it typically takes place against the backdrop of a weaker economy, making it harder for the country to regain market access. It takes place against the backdrop of heavier debts, now including debt to the IMF that is effectively senior to the private debt being restructured (because the IMF generally gets paid back in full), requiring more severe principal reductions ("haircuts") for other creditors. It takes place after private investors have exited, avoiding losses, and thus is a source of moral hazard in encouraging problematic sovereign loans to be made in the first place.[10] For all

[9] These issues arose in the case of Greece in 2010. Still, the vast majority of retrospective assessments, including the IMF's own (IMF 2013a), agree that it would have been better for France and Germany, if they were worried about the impact of a Greek restructuring on their financial systems, to inject capital into their banks and for the institutions of the European Union to support other members threatened with loss of market access, rather than blocking that restructuring.

[10] In all these respects, there is an analogy with the dilemma of a central bank confronted with a potentially insolvent commercial bank in a country lacking a proper resolution regime for insolvent financial institutions (Shafik 2015). Lacking alternatives, there too it may see no alternative to lending in the hope that good news turns up.

Table 2
Countries in an IMF-supported Program with External Debt Restructuring

In IMF-supported programs, 1995–2012	*IMF program*	*External debt restructuring*	*Private creditor bailout?*	*IMF program amount agreed (millions of US dollars)*	*IMF program type*
Pakistan	October 1997	November 1999	Partial	1,552.0	Extended Credit Facility
Russia	July 1998	July 1998	Partial	5,339.8	Supplemental Reserve Facility
Ukraine	September 1998	September 1998	No	2,597.1	Extended Fund Facility
Ecuador	January 2000	July 2000	No	311.8	Structural Adjustment
Argentina	March 2000	April 2005	Yes	23,288.3	Supplemental Reserve Facility
Turkey	December 2000	N/A	Yes	7,523.7	Supplemental Reserve Facility
Uruguay	March 2002	May 2003	No	2,479.5	Supplemental Reserve Facility
Dominican Republic	January 2005	May 2005	No	665.7	Structural Adjustment
Seychelles	November 2008	February 2010	No	25.9	Structural Adjustment
Jamaica	February 2010	February 2010	—	1,267.6	Structural Adjustment
Greece	May 2010	March 2012	Yes	39,341.8	Structural Adjustment
St. Kitts & Nevis	July 2011	April 2012	No	84.5	Structural Adjustment
Jamaica	May 2013	February 2013	No	932.3	Extended Credit Facility
Ukraine	March 2015	August 2015	Yes	17,500.0	Extended Credit Facility

Source: IMF (2014b, p. 13), with authors' updates.
Notes: This table lists countries with an IMF-supported program that have restructured their debt since 1995. Turkey completed a debt exchange in 2001 which the IMF did not categorize as an external debt restructuring. Jamaica's 2013 debt exchange was voluntary although strongly encouraged by the IMF, and it was known that a default was very likely without the operation.

these reasons, "debt restructurings have often been too little and too late," as the IMF (2013b, p. 1) has acknowledged.

How might this tendency be corrected? One possibility is for the IMF to adopt rules preventing it from lending to countries with debts of dubious sustainability. This step would require that there be a restructuring sufficient to restore sustainability in conjunction with the initial IMF loan. The IMF moved in this direction in 2002 following an unsatisfactory experience in Argentina, to which it lent just prior to that country's default. The 2002 "Framework on Exceptional Access" committed the Fund to provide large-scale financing without also requiring a debt restructuring

only if it determined that a country's debt was "sustainable with a high probability" (IMF 2002).[11]

But tying one's hands leaves open the possibility of untying them. In 2010, responding to worries the shock from a Greek restructuring might spill over to other European countries, the IMF Board, on which European countries are heavily represented, invented a "systemic exemption" to the exceptional access framework, permitting the institution to provide Greece with a large financing package despite doubts that the country's debts were sustainable with high probability. The systemic exemption authorized exceptional access in cases where the Board agreed that there was a high risk of "international systemic spillovers." So much, then, for tying one's hands.

The Greek loan was the largest in IMF history. Emerging markets that had been subject to the earlier exceptional access policy complained, not without reason, that large shareholders in the Fund received preferential treatment.

In response, the IMF tried again in 2013–14 to enunciate a rule for debt restructuring. Instead of placing countries into two groups—those whose debts were sustainable with high probability and those that were not—there would now be a third group made up of countries whose debts might be sustainable but not with high probability. The Fund announced that it would only grant exceptional access to countries in the "unsustainable" category if they first engaged in an up-front debt restructuring. In contrast, countries in the new intermediate category might qualify for funding only if the government undertook a "soft re-profiling" of its debt, which meant a bond exchange that involved an extension of debt maturities without any reduction of principal and interest. In conjunction with this change, the systemic exemption would be eliminated.

Why this approach should work better is unclear. The next time fears of contagion arise, nothing will prevent the Board from reviving the systemic exemption. Moreover, requiring a larger share of members, including those in the new intermediate category, to engage in some form of restructuring, even in the form of a soft re-profiling, might encourage investors to rush for the exits at the first sign of trouble. The incentive to minimize the size of the intermediate category, and to delay, would remain.

A different approach to encouraging earlier debt restructuring is to make it less disruptive. To this end, the IMF has encouraged governments to add collective action clauses (CACs) to their sovereign bond contracts. CACs allow a qualified majority of holders of a sovereign bond issue, when they vote to accept a restructuring offer, to impose the same terms on the dissenting minority. These clauses thereby seek to reduce coordination, holdout, and litigation problems that delay

[11] In this context, "exceptional" means access to more than three times a country's nominal IMF quota, which is what is permitted by the Fund's standard procedures. In today's era of large international financial markets, with whose growth IMF quotas have not kept pace, loans larger than three times quota have frequently been required.

restructurings. CACs have been incorporated into the bond contracts of emerging market countries since 2003 and of euro area countries since 2013.

But these collective action clauses apply to individual bond issues, and with the growth of institutional investors it has become easier for a single investor to acquire a sufficiently large share of an issue to block a positive vote. "Super-CACs," where votes are aggregated across all bonds issued by a government, might help, but markets and governments have only limited experience with them.

Starting in 2013, the case for collective action clauses was further complicated by a series of decisions by the District Court for the Southern District of New York, with jurisdiction over Argentine bonds under New York State governing law. The court held that holdout creditors were entitled by the equal treatment (or *pari passu*) clause of their contracts to be paid in full rather than being forced to accept restructured terms. This decision threatened to weaken the incentive for creditors to agree to a restructuring. Again the IMF's response was to promote the use of "enhanced" clauses that eliminate this legal ambiguity by explicitly prohibiting "ratable payments" to holdout creditors (IMF 2015c).

This approach may be effective as a mechanism for mitigating the bias to delay once this contractual language is widely incorporated into new bond issues. But doing so will take time: adding clauses to newly issued bonds will do nothing to dissolve the legal haze covering the inherited stock of long-term, yet-to-mature bonds. This fact has resuscitated arguments for the statutory alternative, where the IMF or another multilateral institution would have the powers of a binding arbitral tribunal—the international equivalent of a national bankruptcy court—to "cram down" restructuring terms on dissenting creditors (Krueger 2002; Stiglitz and Guzman 2015).

Resistance to such proposals is strong. This is especially true of variants that seek to anoint the IMF as international bankruptcy judge or binding arbitrator, because, as Stiglitz and Guzman (2015) put it, the IMF is perceived as "too closely affiliated with creditors."[12] For these reasons, the problem of transactions costs and coordination problems as barriers to timely restructuring, the tendency for the IMF to continue lending for too long, and the resulting moral hazard for investors all remain unresolved. The situation also illustrates how issues of governance and perceptions of impartiality and legitimacy can limit the IMF's own policy room for maneuver.

[12] In addition there is the fact, as Stiglitz and Guzman (2015) go on to note, that the IMF "is a creditor itself." The question to which this points is whether debt to the Fund itself should be subject to restructuring. The IMF has traditionally dismissed this as unacceptable on the grounds that its losses "would mean additional contributions by the international community and some of these countries are in a direr situation than those seeking the delays," as the IMF's managing director Christine Lagarde put it in the context of discussions of a Greek restructuring in 2015. Rodrik (1996) argues that coupling lending with conditionality is a way for a multilateral institution like the World Bank or IMF to create confidence in the quality of its advice—to, in effect, put its money where its mouth is. The presumption that IMF loans are never restructured is problematic from this point of view.

Governance Reform

Problems of governance feed doubts about the IMF's competence and impartiality. Governments and their constituents question whether its advice is well-tailored to their circumstances or simply reflects the self-interest of the institution's dominant shareholders. They point to pressure from the United States for countries to accelerate public-enterprise privatization in the 1990s and pressure from large European countries to avoid a Greek debt restructuring in 2010. IMF members who see themselves as inadequately represented therefore dismiss the Fund's decisions and advice as illegitimate. Their skepticism weakens the role of the Fund as a venue for governments, with guidance from staff and management, to discuss issues with an eye toward reaching a consensual diagnosis of their nature and appropriate treatment. It weakens the role of the Fund as a mechanism for orchestrating the adjustment of national policies with the goal of internalizing cross-border spillovers. Governments oppose giving the IMF the autonomy to engage in "blunt truth-telling" because they doubt that what it tells will be equally truthful in all cases. They resist giving the Fund additional responsibility as the organizer of a sovereign debt restructuring mechanism because they see the institution as overly influenced by a small set of advanced creditor countries.

Hence governance reform that gives appropriate voice and weight in decision making to all members is critical for enhancing the legitimacy and effectiveness of the institution. To understand what this means in practice, it is necessary to explain how IMF governance works. When a country joins the IMF, it must make a certain amount of resources available. These resources are known as its quota. The current quota formula is a weighted average of GDP (weight of 50 percent), openness (30 percent), economic variability (15 percent), and international reserves (5 percent). Voting power in the Fund is largely based on quotas: each country gets a "basic" vote, but in addition it gets more votes in proportion to the size of its quota.

Major decisions at the IMF require an 85 percent majority of voting power. Under the existing system, the United States holds 16 percent of votes, giving it a veto. European Union member states hold about one-third of quota shares, which means that these countries, voting together, have a veto as well. The Managing Director of the IMF has always been European and the First Deputy Managing Director has always been American, a convention that is not unrelated to the distribution of voting power. Quota shares have not been altered over time to reflect fast growth in emerging markets and developing countries, leading these members to be underrepresented in IMF decision-making.

While the ultimate decision-making body in the IMF is the Board of Governors, on which every member country is represented, day-to-day management occurs through the 24-member Executive Board chaired by the Managing Director. Countries with more voting power have their own representatives on the Executive Board: for example, the countries with the five largest quotas—the United States, Japan, France, Germany, and the United Kingdom—each have an individual representative. China, Russia, and Saudi Arabia also have individual representatives, but other

countries share a representative. Decisions in the Board are "by consensus"—votes are rare, in other words. But having a seat at the table still matters; in practice the chairman announces the existence of a consensus when, according to the secretary's tally, a majority of directors have spoken in favor of a motion. Moreover, countries with their own representative have the loudest voice, since that representative is able to channel the views of his or her national government, while representatives of multi-member constituencies must balance the opinions of different governments whose views are not always aligned.

In 2010, the Board of Governors recommended a set of governance reforms, pending ratification by governments. The package included a doubling of quotas, thereby increasing the Fund's lending power.[13] The reforms would realign quota shares so that the ten largest members would be the United States, Japan, France, Germany, Italy, the United Kingdom, Brazil, China, India, and the Russian Federation. It would abolish the right of the five largest quota-holders to appoint their own members of the Executive Board; instead, it provided that all Executive Board members would be elected by groups of countries.

Each reform has costs or disadvantages for certain members. Doubling quotas will require all countries to contribute more. Realignment will reduce Europe's voting share. The right of the United States, United Kingdom, France, Japan, and Germany to replace or dismiss their own Executive Directors would be replaced by a requirement for these countries to form constituencies, whose members would elect a director who would then have greater independence from individual governments.

Although countries with about 80 percent of current voting shares have agreed, implementation of the reform requires an 85 percent supermajority, and therefore it requires the consent of the US Congress—which has been reluctant to ratify the agreement. Disaffected countries have consequently looked elsewhere. Emerging market economies have accumulated vast quantities of international reserves, at considerable cost to themselves (Rodrik 2006), to limit the likelihood that they might need to resort to the IMF for assistance. But self-insurance is costly, since returns on reserve assets are significantly less than those paid on domestic bonds. It is especially costly when reserves can't be used without sending an adverse signal about a country's relative economic strength (Shafik 2015).

These observations point to obvious advantages of pooling reserves and coordinating their use. To this end, various central banks have negotiated bilateral currency swap lines to provide one another with additional insurance. Brazil, Russia, India, China, and Republic of South Africa (the so-called BRICS countries) have signed a Contingent Reserve Arrangement establishing a network of such swaps. The Chiang Mai Initiative Multilateralization (CMIM) of the ASEAN+3

[13] Since 1990, only one increase has been approved (45 percent in 1998), with three other reviews resulting in no increase. The previous largest-ever increase in IMF capital quotas was 60.7 percent in 1958/59. In addition, there was a temporary increase in IMF resources in 2009, at the height of the global financial crisis, arranged via the New Arrangements to Borrow through which 38 members stand ready to provide the Fund with additional financial resources.

countries (the Association of Southeast Asian Nations plus China, South Korea, and Japan) has consolidated the bilateral swap lines of its members in a first step toward pooling their reserves. An Asian Macroeconomic Research Office (AMRO) has been established in conjunction with CMIM with an eye toward developing a surveillance function and providing policy advice. One can begin to see the outlines of alternatives to the IMF developing alongside the institution.

But these new institutions have no track record. They are limited to subsets of countries or regions, and not obviously suited for addressing global problems. To the extent that disturbances affect a particular class of economies (all emerging markets, or all Asian countries, for example), they are not well designed for sharing risks, since reserve pooling buys the participants little in this instance. They are not free of governance problems of their own. These are arguments for strengthening the effectiveness and, to that end, the legitimacy of the IMF.

It is clear why the 2010 governance reforms are supported by countries whose voice and votes in the IMF would be enhanced by them. The argument for why the United States should favor them—despite the fact that it could eventually see its voice and voting share decline—is that the United States benefits from an institution with the legitimacy to act as trusted advisor and emergency lender to governments, to serve as a global venue for discussions of risks to economic and financial stability, and to encourage policy adjustments that take into account cross-border spillovers. Governance reform is necessary in order for the IMF to possess that legitimacy. US Treasury Secretary Jack Lew (2015) has called the reforms "essential to modernizing the IMF's governance and bolstering its ability to respond to urgent international crises."

Governance reform will not automatically make the IMF fit for the future. But without governance reform, the IMF has no future.

References

Beetham, David. 1991. *The Legitimation of Power.* Houndsmill, Basingstoke: Macmillan Education.

Blustein, Paul. 2015. "Laid Low: The IMF, the Euro Zone and the First Rescue of Greece." CIGI Paper no. 61, Center for International Governance Innovation, Waterloo, Canada, April.

Callaghan, Mike. 2014. "2014 Triennial Surveillance Review—External Study—Even-handedness of Fund Surveillance." July 30. Washington, DC: IMF.

Claro, Sebastian, and Claudio Soto. 2013. "Exchange Rate Policy and Exchange Rate Interventions: The Chilean Experience." In *Market Volatility and Foreign Exchange Intervention in EMEs: What Has Changed?* Bank for International Settlements Paper no. 73 (October).

Dahl, Robert A. 1994. "A Democratic Dilemma: System Effectiveness versus Citizen Participation." *Political Science Quarterly* 109(1): 23–34.

de Resende, Carlos. 2014. "An Assessment of IMF Medium-Term Forecasts of GDP Growth." Independent Evaluation Office Background Paper 14/01, International Monetary Fund, Washington, DC.

Dewatripont, Matias, and Gérard Roland. 1992. "The Virtues of Gradualism and Legitimacy in the Transition to a Market Economy." *Economic Journal* 102(411): 291–300.

Drazen, Allan. 2002. "Conditionality and Ownership in IMF Lending: A Political Economy Approach." *IMF Staff Papers* 49: 36–67.

Eichengreen, Barry. 1994. *International Monetary Arrangements for the 21st Century*. Washington, DC: The Brookings Institution.

Eichengreen, Barry. 2003. "Restructuring Sovereign Debt." *Journal of Economic Perspectives* 17(4): 75–98.

Fischer, Stanley. 2001. "Exchange Rate Regimes: Is the Bipolar View Correct?" *Journal of Economic Perspectives* 15(2): 3–24.

Flathman, Richard E. 2012. "Legitimacy." In *A Companion to Contemporary Political Philosophy,* 2nd edition, edited by Robert Goodin, Philip Petit, and Thomas Pogge, volume 2 (of 2 vols.), pp. 678–84. Oxford: Wiley-Blackwell.

Gallagher, Kevin P. 2015. "Contesting the Governance of Capital Flows at the IMF." *Governance* 28(2): 185–98.

Gallagher, Kevin P., and Yuan Tian. 2014. "Regulating Capital Flows in Emerging Markets: The IMF and the Global Financial Crisis." GEGI Working Paper 5, Global Economic Governance Initiative, Boston University, May.

Ghosh, Atish, Timothy Lane, Marianne Schultze-Gattas, Ales Bulír, Javier Hauman, and Alex Mourmouras. 2002. *IMF-Supported Programs in Capital Account Crises*, IMF Occasional Paper 210, September. Washington, DC: IMF.

Global Exchange. 2015. "Top Ten Reasons to Oppose the IMF." San Francisco: Global Exchange.

Grabel, Ilene. 2015. "The Rebranding of Capital Controls in an Era of Productive Incoherence." *Review of International Political Economy* 22(1): 7–43.

Group of Twenty (G-20). 2011. "G20 Coherent Conclusions for the Management of Capital Flows Drawing on Country Experiences." October 15. Washington, DC: Group of Twenty.

Hurd, Ian. 1999. "Legitimacy and Authority in International Politics." *International Organization* 53(2): 379–408.

Independent Evaluation Office. 2006. "The IMF's Advice on Exchange Rate Policy." Issues Paper for an Evaluation by the Independent Evaluation Office (IEO), June, International Monetary Fund, Washington, DC.

Independent Evaluation Office. 2007. "IMF Exchange Rate Policy Advice: Findings and Recommendations." Evaluation Report. Washington, DC: IMF.

Independent Evaluation Office. 2013. "The Role of the IMF as Trusted Advisor." Evaluation Report. Washington, DC: IMF.

Independent Evaluation Office. 2014. "Recurring Issues from a Decade of Evaluation: Lessons for the IMF." Evaluation Report. Washington, DC: IMF.

International Financial Institution Advisory Commission. 2000. *Report to US Congress on Reform of the Development Banks and the International Finance Regime*. Washington, DC: International Financial Institution Advisory Commission.

International Monetary Fund. 2002. "Access Policy in Capital Account Crises," July 29. Washington, DC: IMF.

International Monetary Fund. 2003. "Reviewing the Process for Sovereign Debt Restructuring within the Existing Legal Framework, " August 1. Washington, DC: IMF.

International Monetary Fund. 2006. *Ireland: Financial System Stability Assessment Update*. Staff Country Reports. IMF.

International Monetary Fund. 2007. "IMF Executive Board Adopts New Decision on Bilateral Surveillance over Members' Policies." Public Information Notice No. 07/69, June 21. https://www.imf.org/external/np/sec/pn/2007/pn0769.htm.

International Monetary Fund. 2009. "Greece: 2009 Article IV Consultation Concluding Statement of the Mission." Athens, May 25.

International Monetary Fund. 2012. "The Liberalization and Management of Capital Flows: An Institutional View," November 14. Washington, DC: IMF. http://www.imf.org/external/np/pp/eng/2012/111412.pdf.

International Monetary Fund. 2013a. "Greece: Ex Post Evaluation of Exceptional Access under the 2010 Stand-By Arrangement." IMF Country Report 13/156, June.

International Monetary Fund. 2013b. "Sovereign Debt Restructuring—Recent Developments and Implications for the Fund's Legal and Policy Framework," April 26. Washington, DC: IMF.

International Monetary Fund. 2014a. "2014 Triennial Surveillance Review—Overview Paper," July 30. Washington, DC: IMF.

International Monetary Fund. 2014b. "The Fund's Lending Framework and Sovereign Debt—Annexes," June. Washington, DC: IMF.

International Monetary Fund. 2014c. "Proposed New Grouping in WEO Country Classifications: Low-Income Developing Countries." June, http://www.imf.org/external/np/pp/eng/2014/060314.pdf.

International Monetary Fund. 2015a. "IMF Surveillance." Factsheet, April 13. http://www.imf.org/external/np/exr/facts/surv.htm.

International Monetary Fund. 2015b. "Key

Trends in Implementing the Fund's Transparency Policy," September. Washington, DC: IMF.

International Monetary Fund. 2015c. "IMF Stand-By Arrangement." Factsheet. September 21. https://www.imf.org/external/np/exr/facts/sba.htm.

Johnson, Elizabeth, ed. 1977. *The Collected Writings of John Maynard Keynes,* Vol. 17: *Activities 1920–1922: Treaty Revision and Reconstruction*. Cambridge University Press.

Kose, M. Ayhan, Eswar Prasad, Kenneth Rogoff, and Shang-jin Wei. 2006. "Financial Globalization: A Reappraisal." NBER Working Paper 12484.

Krueger, Anne O. 2002. "A New Approach to Sovereign Debt Restructuring," April. Washington, DC: IMF.

Krugman, Paul. 1998. "The Indispensable I.M.F." *New York Times,* May 15. http://www.nytimes.com/1998/05/15/opinion/the-indispensable-imf.html.

Lew, Jack. 2015. "Remarks of Secretary Lew at the Asia Society of Northern California on the International Economic Architecture and the Importance of Aiming High," March 31. U.S. Department of the Treasury. http://www.treasury.gov/press-center/press-releases/Pages/jl10014.aspx.

Mussa, Michael. 1997. "IMF Surveillance." *American Economic Review* 87(2): 28–31.

Nissan, Sargon. 2015. "As Obituaries Are Written for the World Bank, the IMF is Set to Become Indispensable." *Financial Times, beyondbrics,* May 11.

Orphanides, Athanasios. 2015. "The Euro Area Crisis Five Years After the Original Sin." MIT Sloan Research Paper no. 5147-15, October.

Ostry, Jonathan, Atish Ghosh, Karl Habermeier, Marcos Chamon, Mahvash Qureshi, and Dennis Reinhardt. 2010. "Capital Inflows: The Role of Controls." Staff Position Note 10/04, February. Washington, DC: IMF.

Pisani-Ferry, Jean, Andre Sapir, and Guntram Wolff. 2011. "TSR External Study—An Evaluation of IMF Surveillance of the Euro Area." June 19. Washington, DC: IMF.

Prasad, Eswar S., and Raghuram G. Rajan. 2008. "A Pragmatic Approach to Capital Account Liberalization." *Journal of Economic Perspectives* 22(3): 149–72.

Rodrik, Dani. 1996. "Why Is There Multilateral Lending?" In *Annual World Bank Conference on Development Economics 1995,* edited by Michael Bruno and Boris Plekovic, 167–202. World Bank.

Rodrik, Dani. 2006. "The Social Cost of Foreign Exchange Reserves." *International Economic Journal* 20(3): 253–66.

Roland, Gérard, and Thierry Verdier. 1994. "Privatization in Eastern Europe: Irreversibility and Critical Mass Effects." *Journal of Public Economics* 54(2): 161–83.

Schadler, Susan, María Vinceta Carkovic, Adam Bennett, and Robert Brandon Kahn. 1993. "Recent Experiences with Surges in Capital Inflows." IMF Occasional Paper 108, September.

Scharpf, Fritz W. 1997. "Economic Integration, Democracy and the Welfare State." *Journal of European Public Policy* 4(1): 18–36.

Shafik, Minouche. 2015. "Fixing the Global Financial Safety Net: Lessons from Central Banking." Speech given at the David Hume Institute, Edinburgh, Scotland, September 22. Bank of England. http://www.bankofengland.co.uk/publications/Pages/speeches/2015/841.aspx.

Spence, Michael. 1973. "Job Market Signaling." *Quarterly Journal of Economics* 87(3): 355–74.

Stiglitz, Joseph E., and Martin Guzman. 2015. "A Rule of Law for Sovereign Debt." *Project Syndicate,* June 15. http://www.project-syndicate.org/commentary/sovereign-debt-restructuring-by-joseph-e-stiglitz-and-martin-guzman-2015-06.

Takagi, Shinji, Carlos de Resende, Jerome Prieur, Franz Loyola, and Tam Nguyen. 2014. "A Review of Crisis Management Programs Supported by IMF Stand-By Arrangements, 2008-11." Independent Evaluation Office Background Paper 14/12, International Monetary Fund, Washington, DC.

Tucker, Paul. 2015. "Microprudential Versus Macroprudential Supervision: Functions that Make Sense Only as Part of an Overall Regime for Financial Stability." Paper presented to the Federal Reserve Bank of Boston, October, 2–3.

United States Department of the Treasury. 2000. "Report on IMF Reforms." Report to Congress in accordance with Sections 610 (a) and 613 (a) of the Foreign Operations, Export Financing and Related Programs Appropriations Act, 1999. October20.http://www.treasury.gov/press-center/press-releases/Documents/imfreform.pdf.

Wei, Shang-Jin. 1997. "Gradualism versus Big Bang: Speed and Sustainability of Reforms." *Canadian Journal of Economics* 30(4b): 1234–47.

Journal of Economic Perspectives—Volume 30, Number 1—Winter 2016—Pages 53–76

The New Role for the World Bank

Michael A. Clemens and Michael Kremer

The World Bank was founded to address what we would today call imperfections in international capital markets. Its founders thought that countries would borrow from the Bank temporarily until they grew enough to borrow commercially (NAC 1946, p. 312; Black 1952). Some critiques and analyses of the Bank are based on the assumption that this continues to be its role. For example, some argue that the growth of private capital flows to the developing world has rendered the Bank irrelevant.

However, we will argue that modern analyses should proceed from the premise that the World Bank's central goal is and should be to reduce extreme poverty, and that addressing failures in global capital markets is now of subsidiary importance. The Bank's stated goal is reducing poverty. The overwhelming majority of Bank subsidies from its shareholder countries go to the International Development Association (IDA), its arm for making grants and highly concessional loans to the lowest-income countries. The Bank's greatest impact comes from its role in the dramatic policy changes many developing countries have undertaken in multiple sectors that most economists would consider likely to reduce poverty, either by increasing growth or promoting equity.

Why might donor countries choose to work through an international organization to advance the goal of reducing poverty? Effective aid often involves

■ *Michael Clemens is a Senior Fellow, Center for Global Development, Washington, DC. He is also a Research Fellow of the Institute for the Study of Labor (IZA) in Bonn, Germany. Michael Kremer is the Gates Professor of Developing Societies, Harvard University, and Research Associate, National Bureau of Economic Research, both in Cambridge, Massachusetts.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.1.53 doi=10.1257/jep.30.1.53

negotiating agreements with recipient country governments that include policy reforms. There are economies of scale in negotiating such agreements that can be realized by an entity such as the Bank, and pooling funds into such an entity may also improve donors' collective bargaining position in negotiations with governments. Moreover, we argue that the World Bank's status as a multilateral organization and its technocratic staff enhances its credibility and legitimacy in policy discussions with developing-country governments. This has allowed the Bank to have tremendous policy influence relative to the explicit and implicit subsidies it receives, making it a bargain for those who value its mission of reducing extreme poverty and share its mainstream economic views on what policies best advance that goal.

In this paper, we discuss what the Bank does: how it spends money, how it influences policy, and how it presents its mission. We argue that the role of the Bank is now best understood as facilitating international agreements to reduce poverty, and we examine implications of this perspective. For example, the Bank should not conceptualize its principal activity as capital investment, but instead should consider a broader range of activities and instruments. Moreover, attempts to measure the Bank's success by regressions that use economic growth rates as the dependent variable and disbursements of aid as an explanatory variable will inevitably be quite misleading.

The Cost of the World Bank and Its Focus on Poverty

The World Bank Group operates through divisions with differing roles. Three of these can arguably be interpreted either through the narrow lens of addressing international capital market imperfections or through the broader lens of poverty reduction: the International Bank for Reconstruction and Development (IBRD), which lends to governments; the International Finance Corporation (IFC), which invests in commercial projects; and the Multilateral Investment Guarantee Agency (MIGA), which sells insurance policies to private investors against noncommercial—that is, political—risks. The IBRD, IFC, and MIGA work primarily in middle-income countries, with financial terms ostensibly close to market terms. Their books show them earning profits. However, they receive an implicit subsidy because they have the use of capital that shareholder governments have placed with the Bank, and IBRD shareholder governments also cover the risk that enough deals could go bad that the Bank would need to call on their pledged capital. As discussed below, academics have debated the extent to which those components of the Bank are addressing international capital market failures as opposed to simply providing subsidies.

Other parts of the Bank do not seem focused on addressing capital market failures but are more easily understood as contributing to poverty reduction. These include the Bank's fourth arm, the International Development Association (IDA), and various Trust Funds that the World Bank administers on behalf of donors. IDA was established in 1960. Instead of requiring market interest rates, IDA offers a combination of grants, along with loans on such highly concessional terms that

they amount to grants, for very low-income countries, currently defined as those with per capita income of less than $1,215 per year (at market exchange rates). The World Bank Group also holds Trust Funds supported by donors and used for pre-specified purposes such as fighting AIDS and malaria, immunizing children, or promoting access to education in the developing world. In general, Trust Funds are not intended to generate financial returns.

Table 1 shows a breakdown of the opportunity cost of capital to shareholder governments. The first few rows show the paid-in capital from governments to the IBRD and the IFC, along with the retained earnings that have been accumulated over time from past repayments of loans or liquidating other investments. The first row under "Flows" multiplies the stock of $64.8 billion by 3 percent (approximately based on the 20-year Constant Maturity US Treasury interest rate) to estimate $1.9 billion per year in opportunity cost in the foregone interest from investing in a riskless security. This analysis follows the method of Meltzer (2000) and the Commission on the Role of the MDBs in Emerging Markets (2001), updating their results.

In the second row under "Flows," we estimate the cost of IDA at $8.0 billion, representing annualized donor replenishments over the last decade. The remaining flows describe disbursements from the trust funds. Specifically, we exclude "financial intermediary trust funds," which are pass-through accounts for operations outside the Bank, and include "Bank-executed trust funds," which directly finance Bank operations. The remaining category, "recipient-executed trust funds" (RETF), mixes support for Bank and non-Bank activities. We thus provide estimates with and without RETF disbursements.

The last few rows of Table 1 show an extremely rough estimate for the risk that the Bank might call on its shareholders for additional capital. According to the rules of the World Bank $218.8 billion in capital could be called, although no such call has ever occurred. Current outstanding World Bank loans total $152 billion. As a hypothetical example, if one-third of those loans default completely, representing a loss of $51 billion, $42 billion of this loss could immediately be covered by the existing stock of IRBD capital, leaving $9 billion to be drawn from the callable capital. If the risk to callable capital is equivalent to a 5 percent annual risk of an event such as this, the implicit subsidy paid to cover the risk of such high losses would be about $0.5 billion per year.

Overall, Table 1 estimates that the total subsidy provided by the World Bank Group's shareholders to clients overall is in the range of $11.0–14.2 billion per year, which includes a conservatively large allowance for risk to donors' callable capital. The range in this estimate reflects uncertainty about what portion of recipient-executed trust fund disbursements support operations of the Bank itself. Of this total, aid given through IDA accounts for $8 billion, and most of the $3 billion from recipient-executed trust funds supports activities in IDA-eligible countries (Huq 2010). The vast majority of the donor subsidy goes to very low-income countries. Further details, including our treatment of MIGA and callable capital, are described in the online Appendix available with this paper at http://e-jep.org. To

Table 1
The World Bank Group's Opportunity Cost to Its Shareholders

	US dollars (billions)	*Line*	*Calculation*	*Source*
Stocks				
IBRD				
Paid-in capital	14.0			*a*
Retained earnings	28.3			*a*
Total IBRD	**42.3**	(1)		
IFC				
Paid-in capital	2.5			*b*
Retained earnings	20.0			*b*
Total IFC	**22.5**	(2)		
Total IBRD and IFC	**64.8**	(3)	(1) + (2)	
Flows				
Annual opportunity cost, IBRD+IFC	1.9	(4)	(3) × 3%	
Annual IDA partner grant contribution	8.0	(5)		*c*
Annual trust-fund disbursements				
Bank-Executed Trust Funds (BETF)	0.6	(6)		*d*
Recipient-Executed Trust Funds (RETF)	3.3	(7)		*d*
Annual total cost to donor countries				
without RETF	**10.5**	(8)	(4) + (5) + (6)	
with RETF	**13.7**	(9)	(8) + (7)	
Risk				
Stock: IBRD callable capital (never called)	*218.8*			*a*
Flow: 20-year annual called capital if 1/3 of World Bank loan portfolio never repaid	0.5	(10)		*e*
Total Donor Subsidy				
Annual donor subsidy including risk	**11.0–14.2**		(8, 9) + (10)	

Sources: Sources given in the footnotes.
Notes: Sums of numbers appearing in table may not exactly equal sums shown due to rounding.
[a] World Bank, *Annual Report 2014*, IBRD/IDA Management's Discussion & Analysis and Financial Statements, p. 15.
[b] IFC *Annual Report 2014*, p. 103. Note that the IFC has no callable capital.
[c] Annual average of partner grant contributions in last four IDA replenishments (IDA14 [FY06–08] US$18.0bn; IDA15 [FY09–11] US$25.1bn; IDA16 [FY12–14] US$26.4bn; IDA17 [15–17] US$26.1bn). From the respective IDA Replenishment final reports.
[d] World Bank (2013), *2013 Trust Fund Annual Report*, Concessional Finance and Global Partnerships Vice Presidency, p. 38. We exclude "Financial Intermediary Trust Funds," which pass through the Bank but are not controlled by it.
[e] Current outstanding World Bank loans US$152bn; 1/3 loss means $51bn one-time loss; $42bn of this could be covered with current IBRD paid-in capital and retained earnings; thus $9bn must be drawn from callable capital. Loan maturities range from 8 to 30 years, thus loss spread over approximately 20 years. $9bn/20yrs ≈ $0.5bn/year in ongoing subsidy to offset risk.

be clear, this $11–14 billion is not an estimate of the degree to which the Bank benefits its client countries, nor an estimate of the Bank's cost of capital. Rather, Table 1 estimates the opportunity cost to the Bank's shareholders that arises from the existence of all parts of the Bank.

The World Bank's Policy Role

The annual donor subsidy embodied by the World Bank, on the order of $11–14 billion, is trivial relative to the economies and budgets of recipients and donors. Focusing on these financial flows misses a key part of what the Bank does, for good or ill, in shaping developing-country and donor policy. Here are a few examples.

Agriculture. In the past, many African farmers could only sell to agricultural marketing boards that operated state-run processing facilities and paid a fraction of the world price for export crops. For example, Ghanaian cocoa farmers shortly after independence were only receiving 55 percent of what the board received for selling their produce; Kenyan cotton farmers in the mid-1970s were getting only 48 percent (Bates 1981, pp. 137–140). This practice was common: nine of the ten African countries examined by Kherallah, Delgado, Gabre-Madhi, Minot, and Johnson (2002) had created agricultural marketing boards. The state also ran markets for inputs, such as fertilizer, often delivering them not at all, or only to politically connected farmers, or too late for planting (Lundberg 2005; Lele and Christiansen 1989). The Bank promoted liberalization of agriculture and these monopsonistic agricultural marketing boards are now mostly gone. Seven of the nine countries surveyed by Kherallah et al. (2002) have removed or reformed their state marketing boards.

Health. The World Bank promoted a shift in budgets away from tertiary-care hospitals in capital cities towards community health centers and rural clinics providing basic primary care (Ekman 2004)—for example through Ethiopia's Health Extension Program and Brazil's Family Health Extension Program. Health budgets are now substantially more oriented toward primary care. At one point the Bank pushed for charging fees to at least certain categories of patients, although it has now backed away from this (World Bank 2015a). It now frequently promotes the adoption of pay-for-performance programs within government health services.

Education. The number of out-of-school children and adolescents worldwide fell from 196 million to 124 million between 2000 and 2013 (UNESCO 2015) despite population growth over the period. The Bank has been an important part of the movement for universal primary education, and now that the vast majority of primary-school-age children in the developing world are enrolled in school, the Bank is shifting its focus to improving learning.

Social protection. The Bank has played an important role in the spread of "conditional cash transfer programs"—in which cash transfers to low-income households are linked to children attending school or seeing health care providers. After promising results from Mexico's PROGRESA in the 1990s (now referred to as

Oportunidades) and Brazil's Bolsa Alimentação program, the Bank now supports conditional cash transfer programs in 26 countries (World Bank, undated). The Bank both financially supported national programs and vigorously promoted conditional cash transfer programs, including at international conferences convened for that purpose in Mexico in 2002, Brazil in 2004, and Turkey in 2006. The Bank's researchers also played an important role in rigorously evaluating the impact of these programs, a factor in their rapid diffusion. Such programs have been found to reduce poverty and improve child health and education (Fiszbein and Shady 2009, p. 14).

Regulatory policy. The World Bank's *Doing Business* reports, which provide objective and internationally comparable measures of how different countries regulate the private sector, have been very influential in motivating countries to reduce regulatory barriers to establishing new firms (in this journal, Besley 2015).

Tax policy. While the International Monetary Fund has played a bigger role, the World Bank has supported the dramatic worldwide shift to value-added taxes, which have replaced other taxes widely considered less efficient. Since 1960, a VAT has been adopted as the main consumption tax in over 140 countries (Norregard and Khan 2007; Cnossen 1998).

Trade policy. The World Bank, along with the IMF, has supported shifts from rigid import quotas to more flexible tariffs, along with reductions in tariffs and movements toward "unified" exchange rates in which the same exchange rates are applied to all types of trade. From the 1980s to 1990s, most World Bank adjustment operations were made conditional on trade liberalization (Edwards 1997). Tariffs and statutory barriers to business creation have declined dramatically. In India, for example, the weighted tariff rate has fallen from 54 to 7 percent between 1990 and 2009 (World Bank 2015b).

Conflict recovery. In post-conflict situations, the Bank has supported community-driven development programs and procedures for demobilizing and providing transitional support to ex-combatants, for example, in Bosnia, Cambodia, El Salvador, Lebanon, and Uganda (Kreimer et al. 1998, p. 28).

Property rights. Since the 1960s, the Bank has supported land demarcation and titling programs in Armenia, Bolivia, Guatemala, Indonesia, Malawi, and elsewhere across the developing world (World Bank 2011). Thailand used World Bank support to partition and distribute land to rural residents (Bowman 2004). Whereas governments of developing countries once regularly appropriated private assets, they are now more likely to privatize state assets.

This list of policy areas where developing countries have dramatically changed policies following Bank involvement suggests that an important way to judge the Bank is a) the extent to which the Bank played a causal role in these changes and b) whether these policy changes were appropriate and effective in reducing poverty. We will discuss some of this evidence below, although it is not the main focus of this paper. Indeed, we do not agree with all of these policies or believe they were all well implemented, but we do agree with the general thrust of most of them and believe that they reflect mainstream views within the economics profession.

Assessing the extent to which World Bank actions played a role in the policy changes listed above is of course difficult. Global changes in international relations and in ideology have surely played a role. But there is reason to believe that the Bank accelerated the changes above. Take the spread of conditional cash transfers. While these programs were a Brazilian and Mexican innovation, early World Bank evaluations and promotion of conditional cash transfer policies gave them credibility and legitimacy before a worldwide audience (Fiszbein and Schady 2009; Borges Sugiyama 2011; García and Moore 2012; Ancelovici and Jenson 2013). The Bank also directly financed and helped design such programs in Colombia, Jamaica, El Salvador, Panama, Guatemala, Chile, Senegal, Eritrea, Burkina Faso, Cape Verde, Democratic Republic of the Congo, Ethiopia, and others (Handa and Davis 2006; García and Moore 2012; Pena 2014).

Gavin and Rodrik (1995) conclude that "the Bank is the single most important external source of ideas and advice to developing-country policymakers." This view has empirical support from a survey of 6,731 government officials and development practitioners, from 126 low- and lower-middle-income countries, who were asked for subjective ratings of 57 different aid agencies (Custer, Rice, Masaki, Latourell, and Parks 2015, p. 48). The World Bank ranked first out of 57 for "agenda-setting influence" and fifth for "usefulness of advice." The Bank's lending operations undoubtedly play an important role in its influence. But in our view, the Bank's influence is much larger than what the $11–14 billion effective subsidy, taken in isolation, might be expected to purchase.

The World Bank magnifies its policy influence through a variety of mechanisms. At the national level, the World Bank chairs groups of bilateral donors that negotiate with recipient-country governments. When the Bank withdraws support from a particular government ministry, other donors often follow. This pattern gives the Bank considerable power to influence overall donor flows, and additional leverage in negotiating with client governments. Beyond this, the Bank plays an important role in how ideas move into policy by collecting data, generating ideas, bringing ideas to a wider policy audience, and turning the ideas into specific policies. At the global level, the Bank sets agendas with publications such as the *Doing Business* reports and the annual *World Development Report*. Many national leaders, including Narendra Modi in India and Vladimir Putin in Russia, have explicitly aimed to improve their standings in the *Doing Business* reports by removing barriers to business investment (Besley 2015).

The World Bank's influence derives in large part from its "soft power," by providing a context and a venue for independent policy discussion. In many of the countries where the Bank operates, political competition is focused on patronage or ethnic and cultural issues rather than economic policy. Think tanks are scarce, and the senior civil service is stretched thin. In this environment, Bank staff can have tremendous influence. The Bank recruits internationally and pays salaries higher than those in most civil services, allowing it to attract staff with very strong qualifications. Staffers come from a variety of countries, including developing countries, and are not seen as promoting a single national perspective. Bank staff have access

to policymakers and often build relationships of trust with key civil servants. Politicians will make the overall decision about whether, say, a conditional cash transfer policy should be implemented. But the World Bank can then have huge influence on decisions regarding the details and implementation of the program. When civil servants debate policy options, those who can argue that their preferred position complies with international norms may be less likely to be seen as merely advocating a position designed to advance their personal and bureaucratic interest (Mukand and Rodrik 2005).

A main channel of World Bank influence, and a measure of the Bank's prestige, is the flow of Bank staff to senior policy-making positions in their home countries (Krueger 1998). Frequently mentioned recent examples include: Ngozi Okonjo-Iweala in Nigeria, Luisa Diogo in Mozambique, Montek Ahluwalia in India, Kemal Derviş in Turkey, Richard Webb and Luís Miguel Castilla in Peru, Ellen Johnson Sirleaf and Antoinette Sayeh in Liberia, Moeen Qureshi in Pakistan, Vittorio Corbo in Chile, Ashraf Ghani in Afghanistan, Benno Ndulu in Tanzania, and many others. Former World Bank staffers Suman Bery, Shankar Acharya, and Jayanta Roy played key roles in enacting India's 1991 reforms (Sengupta 2009), contributing to one of the largest reductions in poverty on record.

The Role the World Bank Has Come to Fill

The founding rationale for the World Bank, in sharp contrast to the activities above, was explicitly to address failures in capital markets. But it has been clear for some time that the Bank has evolved beyond that role.

The iconography of the Bank offers clues to what the Bank itself thinks it should do. The motto "Our dream is a world free of poverty" is carved in stone at the Bank's entrance. A three-story-long red HIV ribbon hung on the Washington headquarters of the World Bank in recent years. The bright atrium of the Bank's headquarters features only one monument—not a scale model of an infrastructure project financed by Bank capital, but a commemoration of the Bank's first health project, in which its central role was one of convening and coordination. It is a sculpture honoring the effort to control river blindness, portraying a young African boy leading an older man afflicted by the disease.

The Bank's rhetoric conveys a similar purpose. Four decades ago, then-President Robert McNamara (1973) set the central goal of eradicating absolute poverty, and proceeded to reshape the Bank's operations around interventions that included massive support for smallholder agriculture. The current Bank President Jim Yong Kim describes the dual mission of the Bank as ending extreme poverty by 2030 and boosting prosperity among the poorest 40 percent in low- and middle-income countries.

Providing grants and subsidies to the poorest countries through IDA can clearly be seen as furthering this mission. Even within the IBRD, many loan projects are focused on poverty. For example, 72 percent of current Brazilian IBRD projects are in the relatively impoverished Northeast and North regions, home to

only 36 percent of Brazil's population (World Bank 2015c). Conditional cash transfers are another example of the World Bank's interest in explicitly poverty-focused programs within middle-income countries.

World Bank lending to middle-income countries that can borrow internationally like Brazil seems more difficult to justify based on the original view of the Bank as addressing failures in international capital markets. Bank clients including Brazil, Thailand, Indonesia, and Mexico could easily self-finance all their IBRD loans without meaningfully depleting their international reserves. China provides an extreme example. In mid-2015, China owed the World Bank $13 billion, but held $3.7 trillion in cash reserves. For the many countries that borrow both from the Bank and on purely commercial international capital markets, it is not clear that the Bank is helping improve access to capital. This is true even if the Bank lends at cheap rates due to its superior ability to secure repayment (Bulow and Rogoff 1990). To the extent that Bank loans are senior to commercial loans, countries' ability to borrow from the Bank may simply drive up their cost of borrowing commercially, leaving their overall cost of lending unchanged. This insight has led some researchers to suggest ending most Bank lending to middle-income countries (for example, Bulow and Rogoff 1990, 2005; Meltzer 2003; Einhorn 2006).

However, if one sees the key rationale for the World Bank as addressing extreme poverty, then a different rationale arises for the Bank's continued work in middle-income countries. In 1990, only one-quarter of the world's extreme poor—those living on less than $1 a day at purchasing power parity exchange rates—lived in middle-income countries. Today, with more countries in the middle-income category, about three-quarters of the extreme poor live in middle-income countries (Sumner 2012a). Even as poverty reduction in middle-income countries proceeds in the next decade or two, the majority of the world's extreme poor will continue to be found there (Sumner 2012b). Furthermore, the poor within many middle-income countries are concentrated in specific geographic areas or subnational states or provinces—like certain poor states in India. The Bank could also target sectors of particular importance to poverty reduction, such as social protection.

This rationale for World Bank lending to middle-income countries raises the question of whether such loans actually increase total funds received by the poor. One could easily write down a model in which World Bank lending reduces the amount of its own resources that, say, Brazil transfers to poor areas. Of course, one could also write down a model in which the Bank could negotiate with the Brazilian government and provide lending only if the government also increased spending. This would be an international version of the "flypaper effect" in public finance, in which federal grants fail to fully "crowd out" local-government expenditures. Such effects have certainly been observed in various public-finance settings (in this journal, Hines and Thaler 1995), but research is needed on the extent to which it occurs in the development lending-and-aid setting. Defining the Bank's role in middle-income countries will become more pressing as about half of IDA's member countries are on course within a decade or so to reach the threshold of per capita GDP where they will graduate to IBRD status (Moss and Leo 2011).

The subsidy element in Bank financing is small enough that the transfer element could make only a very modest impact on poverty on its own. An estimated 2.7 billion people live on less than $2 per day (measured using purchasing power parity exchange rates), so if Bank lending had a donor-subsidy element of $11–14 billion, and was perfectly targeted to the poor with no administrative costs, it would increase incomes of the poor by only about $4–5 per year. Again, the Bank's policy role seems much more closely related to its poverty reduction mission than to a narrow focus on addressing international capital market imperfections.

While we have focused on the Bank's articulated mission of reducing extreme poverty, there is also clear evidence that influential World Bank member countries use the institution to advance their interests. From the beginning, the World Bank had a political mission—to use aid to keep countries in the Western political orbit and to compete with the USSR for economic and political influence in third world countries—as well as a narrower economic mission. The political nature of the institution has continued. In particular, the United States has effective veto power over major Bank decisions, and Bank lending tends to follow the commercial and financial interests of the United States (Faini and Grilli 2004; Fleck and Kilby 2006; Kilby 2009). Indeed, US officials explicitly demanded such behavior in recently declassified documents from long ago (McKeown 2009). Also, countries temporarily on the UN Security Council receive more Bank loans (Dreher, Sturm, and Vreeland 2009), and Bank projects may be used to reward countries for General Assembly votes that support priorities of the United States and other high-income countries (Dreher and Sturm 2012). The United States has successfully intervened to limit Bank lending to some countries, including Iran.

In addition, the Bank's choice of anti-poverty policies on which to focus also reflects the interests and domestic politics of key shareholders. For example, the Bank has played a much more active role in arguing that countries should have more open trade policy than in arguing for more open migration policy, a choice that likely arises from reasons of politics more than from reasons of relative effectiveness in reducing global poverty. One of the Bank's technical assistance projects for an agreement on seasonal labor migration from poor South Pacific islands to New Zealand (Luthria and Malaulau 2013) underwent rigorous impact evaluation and was called "among the most effective development policies evaluated to date" (McKenzie and Gibson 2010). But this project has not received much attention, nor has it been followed by major new Bank investments in labor mobility. Of course this may also reflect Bank staff's judgment about the areas where it may have scope to be effective.

Why a Multilateral Organization?

Consider a context in which high-income countries each put some value on certain outcomes in low- and middle-income countries. These low- and middle-income countries also have preferences and make decisions affecting those outcomes. For

example, high-income countries may value alleviating extreme poverty or addressing more specific needs in the developing world, such as education of girls or provision of treatment for those infected with HIV. That value could arise through two separate channels. First, policymakers in high-income countries may believe that conditions in low- and middle-income countries will generate externalities for high-income countries through mechanisms such as terrorism, disease, or migrant flows. Second, policymakers and citizens in high-income countries could directly value certain outcomes in poorer countries. Altruism appears to be an important driver of charitable giving, including in international settings (Cox, Friedman, and Sadiraj 2008; Null 2011; DellaVigna, List, and Malmendier 2012). If Canadian giving reduces infant mortality in Bangladesh, and if Swedes value reductions in infant mortality anywhere in the world, then Canadian giving generates benefits not only for Canada, but also for Bangladesh and for Sweden. Thus theorists from Becker (1974) to Coate (1995) explicitly classify poverty alleviation under altruistic preferences as a public good. The development literature frequently considers poverty alleviation itself as a public good with worldwide reach (Azam and Laffont 2003; Torsvik 2005; Bourguignon and Platteau 2015).

When global public goods are both cause and result of poverty alleviation, multilateral giving can become an efficient choice. For each high-income country acting individually, the optimal rule is to spend on development assistance until the marginal benefit to each particular high-income country of expenditure abroad equals the marginal domestic benefit. Conversely, the low-income country would set the marginal utility of expenditure on the internationally valued good (such as girls' education) equal to that on other goods valued by the developing-country policymaker. But the standard Samuelson (1954) rule suggests that for public goods, it is *collectively* optimal to spend until the total marginal benefit to all countries of expenditure equals the benefit of alternative expenditures. Individual countries have a strong incentive to free-ride on the poverty-alleviation caused by others, resulting in inefficiently little provision of poverty alleviation from the standpoint of each donor individually. This insight suggests large potential gains from coordination.

If global public goods related to poverty alleviation are undersupplied when countries make decentralized decisions, then a global institution like the IDA can address this problem through donor coordination. Contributions are set according to an agreed formula. Once the formula is established, an increase in the contribution from one country, like the United States, is linked to an increase from others. Attempting to achieve this result with a treaty on donor coordination, requiring each nation to follow up individually, would necessarily be an incomplete solution.

Enforcing Mutually Beneficial and Hard-to-Monitor Coordination

Consider first the example of "tied aid," in which countries create policies requiring that part of their aid spending be spent on goods and services supplied by their own firms. For example, in 2009, 67 percent of Greece's foreign aid required the use of Greek contractors (OECD 2011, p. 12). Suppose that when a donor ties aid, a fraction of any aid spending comes back to the donor country but the effectiveness of that aid declines. Donors collectively might be better off joining a pact in

which they promise to untie all their aid. If donors are in symmetric positions, then in equilibrium each donor's firms will get just as much business—they can get hired by other donors no longer tied to their own firms—but the resulting increased efficiency of aid raises utility for all donors. Indeed the 2005 Paris Declaration on Aid Effectiveness (OECD 2005/2008) seeks to reduce tying of aid. However, such an agreement may be hard to enforce. For example, donors may focus on funding sectors in which their own firms are well-placed to bid on contracts. Agricultural exporters like the United States could decide to give aid in the form of food, thus promoting their commercial interest. China could give aid for transport infrastructure that facilitates trade with China. Given these kinds of alternatives, a collective agreement for nations to pass their funding through a multilateral organization may be a more enforceable way to move away from tied aid. Similarly, individual donors acting through bilateral development organizations might want to give preference in hiring to their own citizens. Just as with tied aid, a multilateral system that hires from all countries raises the possibility that (in a symmetrical situation), aid effectiveness would improve and no nationality would necessarily be worse off.

Individual incentives among donors also lead to excessive fragmentation of donor effort. Bilateral donors will engage in "flag-planting"—that is, they will want to have identifiable and specific aid projects of their own that raise the profile of their individual agency, even though such fragmentation of aid efforts comes with an efficiency cost.

Avoiding Duplication and Achieving Economies of Scale

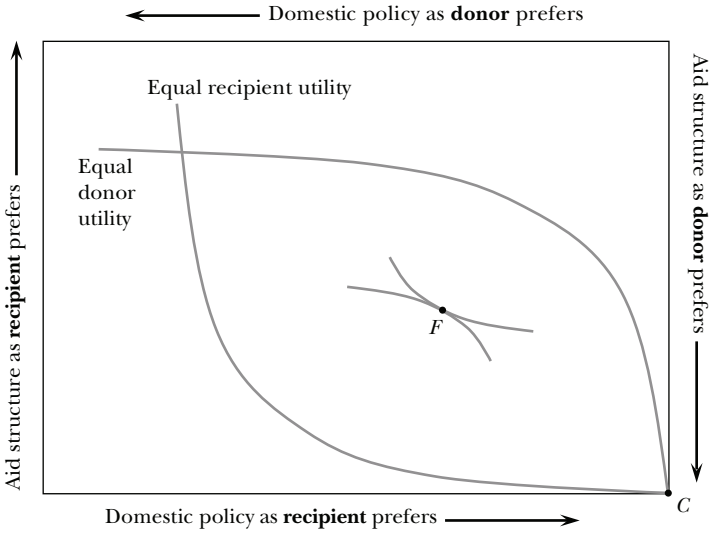
There are important economies of scale and scope in multilateral aid institutions (Sandler 2002; Kanbur 2004; Martens 2005). For example, it would be pointless and wasteful for each individual donor country to undertake its own version of the World Bank's macroeconomic due diligence and poverty-mapping work. It is more efficient for nations to pool their resources and purchase one "aid product" they can all share.

The Bank is able to attract talent in part because it recruits internationally, which typical bilateral development agencies do not do. For any given policy discussion, the World Bank has greater embodied experience in the room—that is, people with personal and professional experience of the issue and place at hand—than most bilateral agencies. Further economies of scale arise for a multilateral aid agency if one takes into account how aid recipient countries must handle accounting so as to comply with the requirements and fiscal years of many different national systems, which can divert substantial resources from other tasks. For example, aid programs in Tanzania are hampered by quarterly accountability reports that the Tanzanian government must submit for over 2,000 donors (Ritzen 2005).

Exploiting Gains from Policy Trade

When donors control some instruments and national governments control others, there will in general be opportunities for gains from negotiations between donor countries and recipient governments to reach the Pareto frontier. Imagine

Figure 1
Pareto-Improving Deal between a Donor and Recipient



Notes: When donors control some instruments and national governments control others, there will in general be opportunities for gains from negotiations between donor countries and recipient governments to reach the Pareto frontier. The potential for gain can be represented in an Edgeworth box. Without negotiation, donor and recipient could start at point *C*, with the donor in full control of aid structure and the recipient in full control of domestic policy. But both parties could achieve higher utility at point *F*, with the donor making concessions on aid structure and the recipient on policy.

that there are two parties, *A* and *B*. Party *A* has full control of the instrument *a,* and party *B* has full control over instrument *b*. Thus, for example, party *A* may be Ghana and instrument *a* might include domestic policies like central bank independence, tariffs, legal requirements for transparency in procurement, free speech, girls' education, transfers to the extreme poor, employment for youth subject to radicalization, and the content of the curriculum in teacher training colleges. Party *B* might be Sweden, and instrument *b* is the amount, sector, and financing terms of Swedish aid offered to Ghana. Both parties have preferences over both instruments.

The potential for gain can be represented in the Edgeworth box of Figure 1. Moving left from the upper-right origin means that domestic policy more closely follows donor preferences, while moving right from the lower-left origin means that domestic policy more closely follows recipient preferences. Moving down from the upper-right origin means that aid structure more closely matches donor preferences, while moving up from the lower-left origin means that aid structure more closely matches recipient preferences. Donor and recipient make tradeoffs according to the indifference curves shown. Without negotiation, donor and recipient could start at point *C*, with the donor in full control of aid structure and the recipient in full control of domestic policy. But both parties could achieve higher utility at point *F*, with the donor making concessions on aid structure and the recipient on policy.

This claim of potential gains does not rest on paternalism—that is, on a claim that the preferences of donors are superior to those of recipients. For example, the government of India presumably cares more about the welfare of the representative citizen in its states of Punjab and in Bihar than does Germany, but Germany likely puts greater relative weight on a dollar of extra GDP for (poorer) Bihar than for (richer) Punjab, than does India. This does not mean that Germany's preferences are superior; it is natural and legitimate for India to value the welfare of its citizens in Punjab. Rather, the opportunity for Pareto improvement arises from the combination of divergent, legitimate preferences and different control over different instruments—such as policies that would help those in Bihar more, or less, than those in Punjab.

Reducing Negotiation Costs of Policy Trade

A deal could potentially be struck at any point between the curves in Figure 1. What point within that area is chosen (or even whether the parties reach an agreement) will depend in part on the relative bargaining strengths of the two parties and how efficiently they bargain. Reaching an efficient bargain requires work. Ideally, negotiations between (say) Ghana and Sweden would look at every policy instrument under Ghana's control and figure out which potential policy changes are most valuable to Sweden and lowest cost for the Ghanaian government to change. Ideal negotiations would simultaneously look at how the amount and composition of Swedish aid to Ghana could most effectively be tweaked to make it more attractive to the Ghanaian authorities while diminishing its appeal to Swedes by as little as possible. The prospect of costly and time-consuming negotiations along these lines for every bilateral aid donor giving to each recipient suggests that it can be efficient for donors to create an institution with shared governance among the donors, such as the World Bank, and to have that entity negotiate with the recipient country.

A multilateral institution can reduce asymmetries of information, increasing the chance of an efficient bargain. If the donors have information that a change in policy in a direction preferred by the donors might be very low cost or even beneficial from the point of view of the recipient country authorities, they would want to find a way to credibly transmit this information. Suppose, for example, that the donors are trying to persuade an oil exporter to drop domestic gasoline subsidies. Donors might want to produce evidence to make their case, but the recipient government may be reluctant to trust the arguments of an official from a bilateral aid agency. They might be much more willing to trust a Brazilian World Bank official with a PhD from University of California at Berkeley who has been involved in similar reform efforts in eight other countries. They might also be willing to trust a series of World Bank publications that review the literature and discuss the range of policy options (including different kinds of phase-outs, targeting, and offsetting policy changes).

Many pathologies can arise in a decentralized aid setting with many small donors. For example, in a decentralized game, a recipient country government

might respond to donor spending in a sector by cutting back its own spending. For example, Pack and Pack (1993) find an inverse correlation between changes in domestic expenditure on health and education and new donor aid to those sectors in the Dominican Republic. Donors may obtain better outcomes if they collectively negotiate with the national government, with the implicit threat point of taking their aid spending to another country if a government is inflexible. They may well be able to crowd in additional government spending in areas of mutual interest. Bilateral aid agencies may not be able to make a credible commitment to walk away from countries where a donor government has a foreign policy interest.

Dialogue and exploring options may be particularly helpful in reaching the Pareto frontier. Exploration might reveal some forms of expenditure that advance the interests of the government and of multiple donors. For example, the World Bank supported a land titling program in Rwanda and, based on a pilot study, realized that women in common-law marriages were losing land titles (Ali, Deininger, and Goldstein 2014). The Rwandan government changed the program to rectify the problem. This was probably a move to the Pareto frontier for the Rwandan government, for donors who believed property rights were important, and for donors who focused on gender issues—relative to a model in which the government and different donors acted independently.

Another advantage of a multilateral organization is that the World Bank's operating directives specifically prohibit it from attempting to exert political influence, whereas there is strong evidence that bilateral donors do attempt to influence elections (Faye and Niehaus 2012). In other cases, of course, donors might want a multilateral agency to focus on these types of political issues. It might be possible to draw some rule-based distinctions for aid—for example between countries with elected governments and unelected governments, which could at least reduce country-to-country tensions that could otherwise be inflamed.

The Role of Legitimacy

In thinking about the Word Bank's success in influencing the global consensus on development policy, it is also important to think in terms of the concept of legitimacy from political science (Bäckstrand 2006). An advantage of delivering aid through international organizations such as the World Bank (or the World Health Organization) is that staff are less likely to be seen as representing the parochial interests of one particular donor, which gives them more legitimacy with the host country government (Rodrik 1996). Markets in policy advice function very poorly, because there are incentives to tell policymakers what they want to hear and what will bring in more consulting contracts—for example, advising countries on how they can use industrial policy to set up an information technology hub, rather than to promote free trade. There is little reason to believe that outcomes would be better if countries simply hired consulting firms to provide policy advice. That is not to say that there are not advantages of competition in aid provision, as in other areas of the economy (Easterly 2002), or that we have any reason to believe that the current balance between the World Bank and other development-focused

organizations is optimal. We claim only that a fully decentralized market in advice would be problematic.

What Should the World Bank Do, and How Should Its Performance Be Assessed?

Should the World Bank focus exclusively on investment projects, or also support raising current consumption of the poor? When the World Bank was founded, it may have seemed that low-income countries were stuck in poverty indefinitely and that financial flows might help them escape this trap. It is harder to make that argument now. Over the past decade, average annual real GDP per capita growth has been 9.4 percent in China and 6.2 percent in India, and sub-Saharan Africa's long period of decline or stagnation in GDP performance has been replaced by a solid, if modest, 2.0 percent annual growth rate. Poorer countries in the same period have grown faster than richer countries: 3.3 percent annual growth for low-income countries, as defined by the World Bank, compared with 0.8 percent for high-income OECD countries. Some countries may be trapped in conflict or in truly dystopian oppression, but for these countries, a lack of access to capital markets is not their main problem. Indeed, run-of-the-mill bad governance is no longer enough to prevent growth. Current growth patterns may or may not persist, but over long periods the evidence for national-level poverty traps is not strong, as Kraay and McKenzie (2014) discuss in this journal (see also Easterly 2006).

If low- and middle-countries are not stuck in a poverty trap, and if donors care about goals such as ensuring that all children have access to basic education or reducing infant and maternal mortality, then it makes much more sense for donors to conceptualize foreign assistance around poverty alleviation rather than around helping countries escape from poverty traps by addressing imperfections in international capital markets.

Indeed, one can make a case that development assistance should seek to raise current consumption rather than investment. Consider a model in which the optimal path of consumption over time is determined by permanent income (that is, expected income over time) and world interest rates. In the benchmark case of this model, development assistance adds to permanent income, which in turn would increase consumption uniformly in the current and all future periods, with the growth rate of consumption unchanged. However, one can write down versions of models along these lines with different sets of constraints in which aid could lead either to more current consumption, or less. For example, if current consumption is low and credit constraints prevent the country from consuming more in the short run, then the optimal response to receiving aid would be to increase current consumption by more than future consumption, which would imply that the country would experience an immediate boost of consumption following the transfer of aid but slower growth of consumption afterwards. Alternatively, in a version of the model in which both consumption and domestic

investment are constrained initially, aid flows could lead to an immediate jump in consumption and thus slower growth of consumption, but also to stimulating investment and a higher growth of GDP.

However, none of these models imply that the gains from development assistance will make a substantial difference to national growth rates. For example, consider an example in which a country saves 20 percent of its income. In this case, a transfer of 5 percent of GDP through aid will lead to a 1 percent of GDP increase in saving and investment. If the rate of return on that additional investment is 10 percent per year, GDP will be greater in subsequent years by one-tenth of 1 percent. Without exotic feedback mechanisms, there is little reason to expect effects of aid to be large or significant in cross-country regressions that use level or growth of GDP as a dependent variable (Swift 2012). Insofar as World Bank aid is a subset of aid overall, it would be even harder to pick up its effects on GDP.

Our view of the Bank as focused on negotiating with government over policies to reduce extreme poverty has implications for its internal structure. For example, to the extent that the Bank is structured on a country-level basis or donors allocate funding on a country-level basis, the Bank bargaining position vis-à-vis national governments is weakened. Conversely, to the extent that the Bank and/or donors within a country can negotiate separately with various ministries or with various levels of government in a federal structure, donors' bargaining power may be strengthened because ministries and subnational governments may compete for donor support.

This reassessment of the role of the Bank suggests that it may want to undertake additional activities beyond lending to governments. For example, the Bank may be able to most effectively fight poverty by spending on areas where both markets and governments have suboptimal incentives to spend, such as on global public goods. Areas where the Bank is already involved in production of international public goods include its support to the Consultative Group for International Agricultural Research (CGIAR); the creation of the pilot Advance Market Commitment for pneumococcus vaccine; and frameworks for cross-border negotiation of water management (Briscoe 2001) and regional labor mobility (Luthria and Malaulau 2013). Many of the Bank's infrastructure projects through the Global Infrastructure Facility have the potential for broad cross-border spillovers, such as support for fiber-optic Internet cables, mobile phone towers, export-processing zones, and airports (World Bank 2015d). Another promising area of Bank activity lies in its efforts to fight money laundering for organized crime and tax evasion through its Financial Markets Integrity unit. The Bank's provision of data—such as through its Living Standards Measurement Surveys and other survey work and its impact evaluation work—has cross-border benefits. The work of the World Bank's research department is also a global public good. A review of the Bank's research department by 28 leading development economists (Banerjee et al. 2006) found that the department had produced "some outstanding work" that has been "hugely influential on global ideas about development"—though it found visible works "where balance was lost in favor of advocacy."

The Bank could also fund scientific research to develop products the poor need, or innovation in ways to deliver development finance—including by strengthening remittance transmission and exploring opportunities for direct and unconditional cash transfers (Blattman and Niehaus 2014). It could potentially in some cases even make transfers directly to the poor, although we have argued that the Bank will typically get bigger bang for the buck through its influence on policy.

Some have argued for a Bank that is focused on international public goods, and in particular around international public goods that involve externalities other than those that are preference-based—such as focusing on climate change—rather than poverty alleviation. While we agree that a multilateral institution in general may be an appropriate place to address the provision of global public goods, some global public goods clearly have less particular relevance to the poor than others (for example, protection of arctic biodiversity or planning to reduce the risks of an asteroid striking Earth). Those who are focused on poverty alleviation, including people from low-income countries, might resist a redirection of existing Bank resources to global public goods in general as opposed to global public goods of particular importance to the poor, such as research on cassava or sleeping sickness.

If one believes that the primary impact of the Bank comes from its specific investments, then one might reasonably evaluate the Bank by assessing what proportion of its investments yield, for example, a 7 percent annual rate of return. If one believes, as we and many other observers do, that the biggest effects of the World Bank arise through its role in influencing developing country policy, then one's assessment of the overall impact of the Bank will hinge primarily on one's beliefs about the effects of these types of policies. A regular bank hopes to obtain a positive return on the vast majority of its investments. In contrast, the Bank could potentially achieve a high social rate of return through a few big wins. If one has a sufficiently favorable view of the regulatory reforms inspired by the Bank's *Doing Business* work, or its role in replacing patchworks of patronage-ridden social programs with conditional cash transfers, then one might believe that the Bank has paid its way, even if the financial performance of Bank loans overall was mediocre. By the same token, the Bank could also have an impact much more negative than the wasting of its funds if it produced harmful policy change. The "fifty years is enough" protesters who sought to close the Bank in the 1990s clearly felt that the Bank's "neoliberal" agenda is harmful, and some conservatives may not support the idea that governments of high-income countries should tap their citizens to provide aid at all. Others take the view that aid, including aid from the Bank, keeps bad governments in power.

Clearly, if one believes that the Bank has reduced poverty primarily by using its financial resources as part of a bargaining process to promote a few important policy reforms, measuring the Bank's impact becomes difficult. Running a regression with country-level outcomes as the dependent variable and World Bank lending as a key explanatory variable will miss the point if a substantial share of the Bank's impact came from the spread of conditional cash transfers, or reforms inspired by the *Doing Business* reports, or the impact of former Bank officials on Indian economic reforms in the 1990s.

Reasonable people can disagree over the specific policies promoted by the World Bank. Each of us is skeptical about some policy initiatives promoted by the Bank, and even when we agree with the overall thrust of the policy, there were often problems in implementation. The devil is often in the details: corrupt or poorly run privatizations may amount to giveaways of state assets, and charging fees at health clinics might be a false economy. But the thrust of most policies promoted by the World Bank has been in line with the mainstream thinking of the economics profession, and our judgment is that, overall, these changes have promoted both equity and efficiency. To put it another way, few economists would advocate a return to the old ways: the agricultural marketing boards; a large share of health care spending focused on tertiary-care hospitals in the capital city rather than rural clinics; higher barriers to business creation; a sprinkling of obscure fees and taxes rather than a basic value-added tax; and so on.

One could ask whether the Bank's policy role could be disconnected from its financial support and from the legitimacy provided by the Bank's status as an international organization. We agree with many observers' judgment that World Bank financing is essential to the Bank's policy influence, both as a pure carrot and as a signal of credibility (Rodrik 1996; Gilbert, Powell, and Vines 1999; Commission on the Role of the MDBs in Emerging Markets 2001; Banerjee and He 2003). Many of these arguments apply independently of whether such financial support is provided as grants or loans. Our analysis is that the Bank's choice of country-partners should take into account the extent to which it can influence policy. In some cases, such as today's Zimbabwe, country policymakers may be resistant to influence. However, donors may sometimes be able to promote poverty reduction by engaging with difficult regimes such as Myanmar that can be nudged in a good direction. Kenya's Daniel arap Moi is often portrayed as having been the paradigmatic example of a corrupt dictator with whom donors should not have engaged. Yet pressure from donors helped encourage him to accept term limits that eventually led to his peaceful departure from office. Judgment calls are inevitable. In some cases, selecting the margin of influence should proceed ministry by ministry rather than country by country.

Finally, our analysis also suggests that the legitimacy of the World Bank may play a role in its effectiveness in influencing policy. The Bank gains legitimacy from its wide membership. The Bank has taken modest steps in reforming its shareholding structure: IBRD shares controlled by developing and transition countries—which determine voting power in the institution—have risen from 43 percent in 2010 to 47 percent today. Maintaining the Bank's legitimacy will require further steps at the Bank, including breaking with the tradition that only US citizens hold its presidency (Rajan 2008) and continuing to adjust its shareholding structure to reflect the relative importance of countries in the world economy. The Bank also gains legitimacy from the perceived professionalism of the staff, including a professional research department, and this professionalism may be quite important. For example, staff with background from the research department often put together the World Bank's *World Development Report,* provide mission support to other Bank staff, and also generate documents that may be useful in interacting with the clients. The existence of a research department that publishes in journals may also attract staff,

similar to Stern's (2004) finding that the opportunity to publish is important in attracting staff to biotech firms.

We believe that the World Bank is, and should be, primarily focused on poverty reduction rather than addressing capital market imperfections. While it is impossible to quantify the Bank's policy influence in a precise way, our judgment is that Bank donors are getting a tremendous amount of policy influence with their limited funding. This influence comes both through deals that link Bank finance to policy reform and through the Bank's soft power. For this reason, allocating more resources to the Bank would be desirable. Other institutions could potentially play a role similar to that of the Bank, but we believe that the Bank has considerable organizational and reputational capital, along with its remarkable track record of policy influence. We see no case for donors to consider withdrawing their funding from the Bank and reallocating it to regional development institutions.

Conclusion

Development advocates often claim that a one-time infusion of capital can generate "sustainable impact," in the sense that after a one-time infusion of funds, the project or investment will generate a stream of income into the future without additional funding. The notion that to be desirable expenditures must be sustainable is politically seductive: for example, the idea of increasing growth and addressing poverty solely by "teaching a man to fish" has comforting appeal. The idea that a temporary infusion of capital would put countries on the path to sustained growth was important for the creation of the World Bank, and it has a continuing influence in the Bank.

However, this "illusion of sustainability" can distort aid and policy decisions (Kremer and Miguel 2007). For example, it can lead to a lack of support for programs that require an ongoing investment in maintenance or management over time, because the requirement for continued funding means that it is not a one-time investment. At a deeper level, the sustainability argument views the justification for aid solely in terms of facilitating escape from poverty traps. But the rapid growth in low-income countries means it is socially efficient for them to consume more now, at least as long as marginal utility is diminishing in consumption and the welfare of the poor has a weight anywhere above the infinitesimally small.

The justification for the World Bank as an aid institution does not rely on capital market failures, or on whether a temporary infusion of capital will suffice. Instead, it is based on poverty reduction. As a multilateral institution, the Bank is well-placed to facilitate Pareto-improving and politically legitimate deals among donors and between donors and developing-country governments. The World Bank has evolved into a role that rests on economic justifications that are sound and defensible, but are profoundly different from those underlying its original stated role. Assessments of the achievements of the Bank and decisions about the Bank's future should be guided by the role the Bank has evolved to fill, and the ways in which the Bank can fulfill that role.

■ *We thank Nancy Birdsall, Alan Gelb, Rachel Glennerster, Gordon Hanson, Enrico Moretti, and Timothy Taylor for helpful comments and Kevin Xie for research assistance. All opinions and errors are those of the authors alone and do not represent their employers or funders.*

References

Ali, Daniel Ayalew, Klaus Deininger, and Markus Goldstein. 2014. "Environmental and Gender Impacts of Land Tenure Regularization in Africa: Pilot Evidence from Rwanda." *Journal of Development Economics* 110: 262–75.

Ancelovici, Marcos, and Jane Jenson. 2013. "Standardization for Transnational Diffusion: The Case of Truth Commissions and Conditional Cash Transfers." *International Political Sociology* 7(3): 294–312.

Azam, Jean-Paul, and Jean-Jacques Laffont. 2003. "Contracting for Aid." *Journal of Development Economics* 70(1): 25–58.

Bäckstrand, Karin. 2006. "Multi-Stakeholder Partnerships for Sustainable Development: Rethinking Legitimacy, Accountability and Effectiveness." *European Environment* 16(5): 290–306.

Banerjee, Abhibit, Angus Deaton, Nora Lustig, Ken Rogoff, and Edward Hsu. 2006. *An Evaluation of World Bank Research, 1998–2005.* Washington, DC: World Bank.

Banerjee, Abhijit V., and Ruimin He. 2003. "The World Bank of the Future." *American Economic Review* 93(2): 39–44.

Bates, Robert H. 1981. *Markets and States in Tropical Africa: The Political Basis of Agricultural Policies.* Berkeley: University of California Press.

Becker, Gary S. 1974. "A Theory of Social Interactions." *Journal of Political Economy* 82(6): 1063–93.

Besley, Timothy. 2015. "Law, Regulation, and the Business Climate: The Nature and Influence of the World Bank Doing Business Project." *Journal of Economic Perspectives* 29(3): 99–120.

Black, Eugene R. 1952. "The World Bank at Work." *Foreign Affairs* 30(3): 402–411.

Blattman, Christopher, and Paul Niehaus. 2014. "Show Them the Money: Why Giving Cash Helps Alleviate Poverty." *Foreign Affairs* 93(3): 117–126.

Borges Sugiyama, Natasha. 2011. "The Diffusion of Conditional Cash Transfer Programs in the Americas." *Global Social Policy* 11(2–3): 250–78.

Bourguignon, François, and Jean-Philippe Platteau. 2015. "The Hard Challenge of Aid Coordination." *World Development* 69: 86–97.

Bowman, Chakriya. 2004. "Thailand Land Titling Project." From the conference on "Scaling Up Poverty Reduction," Shanghai, May 25–27.

Briscoe, John. 2001. "The World Commission on Dams: Lessons Learned about Setting Global Standards." In *Global Public Policies and Programs: Implications for Financing and Evaluation*, edited by Christopher D. Gerrard, Marco Ferroni, and Ashoka Mody, 83–87. World Bank.

Bulow, Jeremy, and Kenneth Rogoff. 1990. "Cleaning Up Third World Debt without Getting Taken to the Cleaners." *Journal of Economic Perspectives* 4(1): 31–42.

Bulow, Jeremy, and Kenneth Rogoff. 2005. "Grants versus Loans for Development Banks." *American Economic Review* 95(2): 393–97.

Cnossen, Sijjbren. 1998. "Global Trends and Issues in Value Added Taxation." *International Tax and Public Finance* 5(3): 399–428.

Coate, Stephen. 1995. "Altruism, the Samaritan's Dilemma, and Government Transfer Policy." *American Economic Review* 85(1): 46–57.

Commission on the Role of the MDBs in Emerging Markets. 2001. *The Role of the Multilateral Development Banks in Emerging Market Economies: Findings of the Commission on the Role of the MDBs in Emerging Markets.* (The committee was co-chaired by Gurria, José Angel Gurria and Paul Volcker; preparation of the report lead by Nancy Birdsall.) Carnegie Endowment for International Peace.

Cox, James C., Daniel Friedman, and Vjollca Sadiraj. 2008. "Revealed Altruism." *Econometrica* 76(1): 31–69.

Custer, Samantha, Zachary Rice, Takaaki Masaki, Rebecca Latourell, and Bradley Parks. 2015. *Listening to Leaders: Which Development Partners Do They Prefer and Why?* Williamsburg, VA: AidData.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics* 127(1): 1–56.

Dreher, Axel, and Jan-Egbert Sturm. 2012. "Do the IMF and the World Bank Influence Voting in the UN General Assembly?" *Public Choice* 151(1–2): 363–97.

Dreher, Alex, Jan-Egbert Sturm, and James Raymond Vreeland. 2009. "Development Aid and International Politics: Does Membership on the UN Security Council Influence World Bank Decisions?" *Journal of Development Economics* 88(1): 1–18.

Easterly, William. 2002. "The Cartel of Good Intentions: The Problem of Bureaucracy in Foreign Aid." *Journal of Policy Reform* 5(4): 223–50.

Easterly, Willaim. 2006. "Reliving the 1950s: The Big Push, Poverty Traps, and Takeoffs in Economic Development." *Journal of Economic Growth* 11(4): 289–318.

Edwards, Sebastian. 1997. "Trade Liberalization Reforms and the World Bank." *American Economic Review* 87(2): 43–48.

Einhorn, Jessica. 2006. "Reforming the World Bank." *Foreign Affairs* 85(1): 17–22.

Ekman, Björn. 2004. "Community-based Health Insurance in Low-Income Countries: A Systematic Review of the Evidence." *Health Policy and Planning* 19(5): 249–70.

Faini, Riccardo, and Enzo Grilli. 2004. "Who Runs the IFIs?" CEPR Discussion Paper 4666.

Faye, Michael, and Paul Niehaus. 2012. "Political Aid Cycles." *American Economic Review* 102(7): 3516–30.

Fiszbein, Ariel, and Norbert Schady. 2009. *Conditional Cash Transfers: Reducing Present and Future Poverty.* A World Bank Policy Research Report. With Francisco H. G. Ferreira, et al. Washington, DC: World Bank.

Fleck, Robert K., and Christopher Kilby. 2006. "World Bank Independence: A Model and Statistical Analysis of US Influence." *Review of Development Economics* 10(2): 224–40.

García, Marito, and Charity M. T. Moore. 2012. *The Cash Dividend: The Rise of Cash Transfer Programs in Sub-Saharan Africa.* Washington, DC: World Bank.

Gavin, Michael, and Dani Rodrik. 1995. "The World Bank in Historical Perspective." *American Economic Review* 85(2): 329–34.

Gilbert, Christopher, Andrew Powell, and David Vines. 1999. "Positioning the World Bank." *Economic Journal* 109(459): 598–633.

Gilson, Lucy. 1997. "The Lessons of User Fee Experience in Africa." *Health Policy and Planning* 12(3): 273–85.

Handa, Sudhanshu, and Benjamin Davis. 2006. "The Experience of Conditional Cash Transfers in Latin America and the Caribbean." *Development Policy Review* 24(5): 513–36.

Hines, James R., and Richard H. Thaler. 1995. "The Flypaper Effect." *Journal of Economic Perspectives* 9(4): 217–26.

Huq, Wahida. 2010. "Analysis of Recipient Executed Trust Funds." CFP Working Paper Series, no. 5; Report no. 58951, World Bank.

Kanbur, Ravi. 2004. "Cross-Border Externalities and International Public Goods: Implications for Aid Agencies." In *Global Tensions: Challenges and Opportunities in the World Economy*, edited by Lourdes Benería and Savirtri Bisnath. Routledge, pp. 65–75.

Kherallah, Mylene, Christopher L. Delgado, Eleni Z. Gabre-Madhin, Nicholas Minot, and Michael Johnson. 2002. *Reforming Agricultural Markets in Africa: Achievements and Challenges.* International Food Policy Research Institute.

Kilby, Christopher. 2009. "The Political Economy of Conditionality: An Empirical Analysis of World Bank Loan Disbursements." *Journal of Development Economics* 89(1): 51–61.

Kraay, Aart, and David McKenzie. 2014. "Do Poverty Traps Exist? Assessing the Evidence." *Journal of Economic Perspectives* 28(3): 127–48.

Kreimer, Alcira, John Eriksson, Robert Muscat, Margaret Arnold, and Colin Scott. 1998. *The World Bank's Experience with Post-Conflict Reconstruction.* World Bank Publications.

Kremer, Michael, and Edward Miguel. 2007. "The Illusion of Sustainability." *Quarterly Journal of Economics* 122(3): 1007–65.

Krueger, Anne O. 1998. "Whither the World Bank and the IMF?" *Journal of Economic Literature* 36(4): 1983–2020.

Lele, Uuma, and Robert E. Christiansen. 1989. *Markets, Marketing Boards, and Cooperatives in Africa: Issues in Adjustment Policy.* World Bank.

Lundberg, Mattias. 2005. "Agricultural Market Reforms." In *Analyzing the Distributional Impact of Reforms,* vol. 1, edited by Aline Coudouel and Stefano Paternostro, pp. 145–212. World Bank. http://documents.worldbank.org/curated/en/2005/01/6065287/analyzing-distributional-impact-reforms-vol-1-2.

Luthria, Manjula, and Mai Malaulau. 2013. "Bilateral Labor Agreements in the Pacific: A Development-Friendly Case Study." In *Let Workers Move: Using Bilateral Labor Agreements to Increase Trade in Services*, edited by Sebastián Sáez. Washington, DC: World Bank, 129–48.

Martens, Bertin. 2005. "Why Do Aid Agencies Exist?" *Development Policy Review* 23(6): 643–63.

McKenzie, David, and John Gibson. 2010. "The Development Impact of a Best Practice Seasonal Worker Policy." World Bank Policy Research Working Paper 5488.

McKeown, Timothy J. 2009. "How US Decision-makers Assessed their Control of Multilateral Organizations, 1957–1982." *Review of International Organizations* 4(3): 269–91.

McNamara, Robert S. 1973. Address to the Board of Governors, 1973 Annual General Meeting, Nairobi, Kenya, September 24. World Bank.

Meltzer, Allan H. 2000. *Report of the International Financial Institution Advisory Commission*. Washington, DC: Government Printing Office.

Meltzer, Allan H. 2003. In "The Future of the IMF and World Bank: Panel Discussion." *American Economic Review* 93(2): 45–50.

Moss, Todd J., and Benjamin Leo. 2011. "IDA at 65: Heading toward Retirement or a Fragile Lease on Life?" Working Paper 246, Center for Global Development.

Mukand, Sharun W., and Dani Rodrik. 2005. "In Search of the Holy Grail: Policy Convergence, Experimentation, and Economic Performance." *American Economic Review* 95(1): 374–83.

NAC. 1946. "Statement of the Foreign Loan Policy of the United States Government by the National Advisory Council on International Monetary and Financial Problems." NAC Document 70-A, February 21. In *Annual Report of the Secretary of the Treasury on the State of the Finances*, Treasury Dept. Doc. 3146, pp. 300–315. Office of the Secretary of the Treasury. Washington, DC: Government Printing Office.

Norregaard, John, and Tehmina S. Khan. 2007. "Tax Policy: Recent Trends and Coming Challenges." Working Paper no. 7-274, International Monetary Fund.

Null, C. 2011. "Warm Glow, Information, and Inefficient Charitable Giving. *Journal of Public Economics* 95(5): 455–65.

OECD. 2005/2008. "*The Paris Declaration on Aid Effectiveness* and the *Accra Agenda for Action*." (The Paris Declaration on Aid Effectiveness is dated 2005; the Accra Agenda for Action is dated 2008.) Organisation for Economic Co-operation and Development. http://www.oecd.org/dac/effectiveness/34428351.pdf.

OECD. 2011. "Implementing the 2001 DAC Recommendation on Untying Aid: 2010–2011 Review." Development Co-Operation Directorate, Development Assistance Committee DCD/DAC(2011)4/REV1. Organisation for Economic Co-operation and Development. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DCD/DAC(2011)4/REV1&docLanguage=En.

Pack, Howard, and Janet Rothenberg Pack. 1993. "Foreign Aid and the Question of Fungibility." *Review of Economics and Statistics* 75(2): 258–65.

Pena, Paola. 2014. "The Politics of the Diffusion of Conditional Cash Transfers in Latin America." Brooks World Poverty Institute Working Paper 20114, University of Manchester.

Rajan, Raghuram G. 2008. "The Future of the IMF and the World Bank." *American Economic Review* 98(2): 110–115.

Ritzen, Jozef. 2005. *A Chance for the World Bank*. Anthem Press.

Rodrik, Dani. 1996. "Why is There Multilateral Lending?" *Annual Bank Conference on Development Economics 1995*. Washington, DC: World Bank.

Samuelson, Paul A. 1954. "The Pure Theory of Public Expenditure." *Review of Economics and Statistics* 36(4): 387–89.

Sandler, Todd. 2002. "Financing International Public Goods." In *International Public Goods. Incentives, Measurement and Financing*, edited by Marco Ferroni and Ashoka Mody, pp. 81–118. Kluwer Academic Publishers and the World Bank.

Sengupta, Mitu. 2009. "Making the State Change Its Mind—The IMF, the World Bank and the Politics of India's Market Reforms." *New Political Economy* 14(2): 181–210.

Stern, Stern. 2004. "Do Scientists Pay to Be Scientists?" *Management Science* 50(6): 835–53.

Stiglitz, Joseph E. 1999. "The World Bank at the Millennium." *Economic Journal* 109(459): 577–97.

Sumner, Andy. 2012a. "Where Do the Poor Live?" *World Development* 40(5): 865–77.

Sumner, Andy. 2012b. "Where Will the World's Poor Live? An Update on Global Poverty and the New Bottom Billion." CGD Working Paper 305, Center for Global Development.

Swift, Daniel. 2012. "ODA and Growth: A Reality Check." Unpublished paper, US Agency for International Development.

Torsvik, Gaute. 2005. "Foreign Economic Aid; Should Donors Cooperate?" *Journal of Development Economics* 77(2): 503–515.

UNESCO. 2015. "A Growing Number of Children and Adolescents Are Out of School as Aid Fails to Meet the Mark." Policy Paper 22/ Fact Sheet 31. Montréal: UNESCO Institute for Statistics, July. http://unesdoc.unesco.org/images/0023/002336/233610e.pdf.

World Bank. 2011. "Land Tenure Policy: Securing Rights to Reduce Poverty and Promote Rural Growth." Report no. 83199.

World Bank. 2015a. *World Development Report 2015: Mind, Society, and Behavior.* Washington, DC: World Bank.

World Bank. 2015b. "Trade." A webpage. http://data.worldbank.org/topic/trade.

World Bank. 2015c. "Projects – Brazil." Information at a website. http://www.worldbank.org/projects/search?lang=en&searchTerm=&tab=map&countryshortname_exact=Brazil.

World Bank. 2015d. "Global Infrastructure Facility." Webpage. http://www.worldbank.org/en/programs/global-Infrastructure-facility#4.

Journal of Economic Perspectives—Volume 30, Number 1—Winter 2016—Pages 77–94

The World Bank: Why It Is Still Needed and Why It Still Disappoints

Martin Ravallion

Poor people find it harder to save, so poor countries can find it difficult to finance needed investments from domestic savings alone. In an ideal world, this would not be a problem; capital would flow from high-income capital-rich countries to low-income capital-poor countries, because the marginal return should be higher in countries where capital is relatively scarce. But that was not what people saw happening in the world 70 years ago. In the years just after World War II, global capital markets were thin and not trusted as a source of finance. It seemed that new institutions were needed.

In response, delegates from 44 countries met in 1944 at a hotel in Bretton Woods, New Hampshire, and agreed to create the International Monetary Fund (IMF) and the International Bank for Reconstruction and Development (IBRD). The latter is a core component of what came to be known as the World Bank Group, or more often the World Bank. The IMF was charged with managing imbalances of payments to avoid destabilizing currency devaluations, while the World Bank was to be the channel for longer-term development finance.

Much has changed since then. There have been prominent calls for radically reforming the World Bank, or even closing it. Two main concerns have been raised by the Bank's critics. The first is that the Bank's efforts are largely wasted because poor countries face nonfinancial constraints that limit their development. The second is

■ *Martin Ravallion holds the inaugural Edmond D. Villani Chair of Economics, Georgetown University, Washington, DC. Prior to joining Georgetown in December 2012, he worked for the World Bank for 24 years, serving most recently as the director of its research department. His email address is mr1185@georgetown.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.1.77 doi=10.1257/jep.30.1.77

that global financial markets are no longer thin and can now serve the Bank's original role. In 1945, the global stock of international investments (measured by asset values) represented 5 percent of world GDP, while 50 years later it had risen to 62 percent (Obstfeld and Taylor 2004). Today, developing countries turn often to the private sector to finance investment; World Bank lending in 2012 represented only about 5 percent of the aggregate private capital flows to developing countries.

Does the World Bank still have an important role to play? How might it fulfill that role? The paper begins with a brief account of how the Bank works. It then argues that, while the Bank is no longer the primary conduit for capital from high-income to low-income countries, it still has an important role in supplying the public good of development knowledge—a role that is no less pressing today than ever. This argument is not a new one. In 1996, the Bank's President at the time, James D. Wolfensohn (1996), laid out a vision for the "knowledge bank," an implicit counterpoint to what can be called the "lending bank." A knowledge bank might serve a number of functions. It can be a broker that taps into existing knowledge and redirects it to needy clients (which was the role emphasized by Wolfensohn). But this vision is rather limited. There is also the task of identifying pressing knowledge gaps—our key areas of ignorance constraining development—and filling those gaps.

The paper argues that the past rhetoric of the "knowledge bank" has not matched the reality. An institution such as the World Bank—explicitly committed to global poverty reduction—should be more heavily invested in knowing what is needed in its client countries as well as in international coordination. It should be consistently arguing for well-informed pro-poor policies in its member countries, tailored to the needs of each country, even when such policies are unpopular with the powers-that-be. It should also be using its financial weight, combined with its analytic and convening powers, to support global public goods. In all this, there is a continuing role for lending, but it must be driven by knowledge—both in terms of what gets done and how it is geared to learning. The paper argues that the Bank disappoints in these tasks but that it could perform better.

How the World Bank Functions

The World Bank currently has 188 member countries, and it employs over 12,000 staff working from 120 offices globally. The Bank is divided into five groups, which together disbursed $44 billion in 2014. Two of the groups are focused on lending and aid. The International Bank for Reconstruction and Development (IBRD) is the original World Bank institution. It primarily makes loans to middle-income countries and made $19 billion in loans in 2014. The International Development Association (IDA) provides grants and loans at favorable terms targeted to low-income countries; it disbursed $13 billion in 2014. The Bank's cumulative lending (IBRD + IDA) between 1945 and 2011 was $788 billion, spread over about 180 countries of which the largest share went to India (11.3 percent),

followed by Mexico (6.5 percent), Brazil (6.3 percent), China (6.3 percent), and Indonesia (5.5 percent).

The other three members of the World Bank Group are more focused on directly encouraging private-sector activity. The International Finance Corporation (IFC) lends to private institutions and disbursed $9 billion in 2014. (IFC profits also support IDA.) The Multilateral Investment Guarantee Agency (MIGA) focuses on insurance and credit guarantees for private investors. The International Center for the Settlement of Investment Disputes (ICSID) is a forum for disputes between investors and governments; use of the ICSID process is written into many international investment treaties, domestic investment laws, and specific contracts.

The World Bank raises money in various ways. Most of its income comes from lending the Bank's own capital, which includes both funds accumulated over time and funds paid in by the member countries. The Bank can sell AAA-rated bonds in the global financial markets, thanks to its conservative lending policies relative to its capital. It can then re-lend these funds at higher interest rates through the IBRD. For the low-interest loans and grants made through IDA, 40 donor countries contribute funds triennially. The Bank also receives some funds from donors for administering their aid and from client countries for reimbursable services. Finally, there are short-term trust funds, which the Bank manages on behalf of other nonprofit agencies.

Formally, the World Bank is run by a Board of Governors, with representatives from all the member countries, meeting annually. The Executive Board (hereafter "the Board") meets regularly and comprises representatives appointed by the six largest shareholders—currently China, France, Germany, Japan, the United Kingdom, and the United States—plus 19 members each representing groups of countries. Membership entails a minimum weight in voting, which then rises according to ownership of the Bank's capital stock. The Bank President presides over the World Bank Group and chairs the Board. Reforms in 2010 increased the voting rights of borrowing countries, notably (but not only) China, which is now the third-ranked in IBRD votes after the United States and Japan.

The Bank's lending operations have long been organized around country teams, each led by a country manager/director. The countries are assigned to six regional groupings, each with its Vice President. This country-based model is backed up by some cross-cutting central units. For example, there are sectoral support units now called Global Practices, which provide specialized expertise and project lending in agriculture, education, energy, health, the environment, transportation, and other areas. The Development Economics Vice Presidency (DEC) is the chief research arm of the World Bank, led by a Chief Economist who reports directly to the President. There is also an Independent Evaluation Group (IEG), which has the task of evaluating World Bank lending and projects in both the public and private sectors, and reports to the Board.

The World Bank is not the only international development bank. The three largest regional development banks—the Inter-American Development Bank, the Asian Development Bank, and the African Development Bank—have expanded

operations since the 1960s. They are collectively still smaller than the World Bank, but there is clearly a degree of competition (as discussed in Kanbur 2003). Recently, China has taken the lead in establishing a new Asian Infrastructure Investment Bank. The New Development Bank has also been created by the BRICSs: Brazil, Russia, India, China, and South Africa. It appears that expanding private finance is not displacing public development finance globally, but instead both are expanding.

Why the World Bank is Still Needed

While developing countries have greatly improved their access to global capital markets, private capital flows have tended to be selective, not reaching all countries and sectors. The Bank has a role in facilitating private finance when needed. But underdevelopment is not only due to a lack of external finance; it has deeper causes in poor policy-making and governance in developing countries—in short, their political economy.

In making the case for the World Bank today, one cannot simply point to unfunded projects. One must explain how the Bank's lending or aid addresses the reasons why such projects are not already funded. That requires the Bank to be a credible knowledge leader.

Why a Development Bank?

Capital markets encounter persistent problems of uninsured risk (including from asymmetric information), externalities, and contract enforcement. Private sector lending to low-income countries can be risky. While private capital flows have increased substantially, the flows are still quite volatile, with potentially destabilizing macroeconomic effects. The Bank can address these problems in several ways: by making loans directly; by giving the private sector a positive signal through its decision to make loans; and by providing trusted sources of information that give the private sector the ability to assess risk and to make loans. The Bank's ability to develop and disseminate knowledge underpins its ability to fulfill these roles.

Development knowledge has properties of a public good. Agents in the private sector have little obvious incentive for publicly documenting what they have learned about development, so that it can be available for the benefit of others. Scale economies in knowledge production can also entail large costs at the outset. If the supply of development knowledge depends on voluntary contributions by individuals or countries, then there will be too little supply. In principle, an institution such as the Bank is well suited to resolving the deficiencies of decentralized knowledge provision.

Development challenges spillover across country borders in the form of pandemics, wars, refugee migrations, and environmental disasters. A global financing institution can play a role in helping to address these regional and global public bads. For example, during the recent Ebola outbreak in West Africa, the Bank deployed $400 million for improving health systems in the affected countries. In 2015, the Bank

created a Pandemic Financing Facility to provide health workers, equipment, and drugs in response to future pandemics. The discussion will return to these issues.

Why a *World* Bank?

The arguments in favor of multilateral development lending and aid reflect concerns over how national governments politicize aid in either bilateral or regional settings. Too often, country preferences over who receives lending or aid reflect foreign-policy considerations and historical ties rather than genuine need or efficacy. Simulations by Collier and Dollar (2002) suggest that an allocation of aid that minimized aggregate poverty would differ greatly from the allocation that existed in the 1990s. According to their calculations, the poverty-minimizing allocation would have almost doubled poverty reduction relative to the actual allocation. Furthermore, bilateral aid has often been tied to recipient countries buying goods and services produced by the donor, a practice that has reduced the real value of aid (Temple 2010). Evidence also points to bilateral aid being used to buy support in major world forums, such as the UN Security Council (Kuziemko and Werker 2006).

The "pet projects" of national development ministries (possibly serving the interests of a local lobby in the donor country) need not make a lot of sense in the context of a sound strategy for poverty reduction. In contrast, a well-functioning global institution can generate economies of scale in knowledge and lending that are out-of-reach for a bilateral agency or even a regional institution. A global institution can also encourage broader participation by high-income countries, thus reducing what otherwise could be a severe free-rider problem. A multilateral institution can also serve a coordination function, embracing both bilateral and regional development lending and aid programs.

Escaping Traps and Overcoming Constraints

A source of market failure that has been prominent in arguments for development assistance concerns the scope for poverty traps. Rosenstein-Rodan (1943), who went on to be a prominent economist at the World Bank in 1947–53, pointed to complementarities between the investments made by different firms in an underdeveloped economy. If all firms invested, then they would all do well, but no individual firm has the incentive to invest when others do not. Development stalls in the inferior equilibrium—a poverty trap. The idea of coordination failures prompted Rosenstein-Rodan and others to advocate what came to be known as a "big push"—a large injection of aid for low-income countries. More recently, Sachs (2005) invoked the poverty trap idea to argue for an increase in development aid. Better public information can also help address coordination failures stemming from complementarities in the investment decisions of firms (Englmaier and Reisinger 2008).

While the idea of a poverty trap as a low-level attractor has been influential, and there can be little doubt that such inferior equilibria exist, their empirical relevance in normal times is less evident (see, for example, Kraay and McKenzie 2014). Models with multiple equilibria are not easily identified empirically and a slow adjustment processes over time can be mistaken for a trap. For the purposes of the present

argument, however, it is not essential to resolve the question as to whether constraints on development are best viewed as "traps" or as substantial hindrances. Instead, one can postulate the existence of constraints on development that the private sector or bilateral agencies cannot address on their own. Hausmann, Rodrik, and Velasco (2008) provided an influential formulation of the development problem in terms of binding constraints specific to each country. The policy idea here is not necessarily of the "big push" variety, although that may well be valid in some cases. Instead, it is to assess for each country what is constraining poverty reduction and to target policy reforms accordingly. Identifying the relevant constraints is not easy and requires considerable country-level expertise. Relaxing constraints may require complementary public inputs such as spreading technical knowledge, supporting more capable public administrations, and helping to supply public goods.

The Continuing Case for Bundling Lending with Knowledge

There has been an ongoing controversy over the extent to which development assistance has benefited the recipient countries. Some observers have argued that badly governed people in a poor country will be worse off with aid as it will reward and support the regime (for example, Deaton 2013, chap. 7). While the attribution problems are severe, given that aid is endogenous, my own review of the evidence from many studies suggests a credible case that past development aid has helped (Ravallion 2016, chap. 9).

My purpose here is not to revisit this debate, but to point out that the experience with development assistance has lessons to teach about what works and what doesn't. Learning these lessons requires the development institution to be centrally focused on generating and disseminating relevant knowledge at country, regional, and global levels. The gains from bundling knowledge with lending provide the key rationale for the Bank's existence in a world of more developed capital markets (as argued by Gilbert, Powell, and Vines 1999). It also points to a key difference between the World Bank and dedicated research institutions, including academia.

The rest of this paper will argue that a valid case for World Bank lending operations remains, but knowledge must drive that lending—both informing the nature of the lending and learning from it—rather than simply serving lending when it happens to be called upon. From this perspective, the Bank is falling short of its potential.

Why the World Bank Still Disappoints

Sound evaluations both before and after its operations are clearly crucial to a knowledge bank, so this topic is a good place to start. The discussion then turns to other things to be expected of a knowledge bank, and how the World Bank performs.

Evaluations of Lending Operations

The first question we would surely ask of a knowledge bank is whether it establishes a sound prior case for its own interventions and systematically assesses

whether that case turned out to be valid. The World Bank has not, however, lived up to this ideal. Evaluation is generally weak and unbalanced, both before and after implementation. This reflects a lack of focus on the welfare outcomes of projects and policies. Instead of studying the effect on its stated goal of poverty reduction, the focus tends to be on monitoring inputs—for example, schools built rather than education attainments (Gaarder and Bartsch 2015).

A true knowledge bank will address questions like: Why is the proposed project needed? How does it relate to overall development goals? What are the market, or governmental, failures it addresses? What are its distributional goals? What are the trade-offs? Social cost–benefit analysis provides the economic framework for addressing these questions (Devarajan, Squire, and Suthiwart-Narueput 1997). The Bank was once a leader in cost–benefit analysis, but this is no longer true. While the Bank's operational directives call for cost–benefit analysis, it is not implemented for most Bank projects (World Bank 2010). The proportion of projects quoting an expected rate of return has fallen over time.

Cost–benefit analysis has clearly fallen out of favor among World Bank staff and managers. There have been justifiable concerns about the quality of the key inputs to the analysis, notably on the benefits side—both the magnitude of the benefits and their monetary values. Uncertainty about key parameters creates scope for manipulation by project staff keen to get their loan approved. But these are not good reasons for abandoning project appraisal. We still need to know what the case is for the project, given what we know and recognizing the uncertainties. The identified knowledge gaps should then be addressed in follow-up work to reduce the uncertainties for future appraisals. Checks can be done on the quality of the analysis.

The decline in cost–benefit analysis at the World Bank came with a welcome rise in the use of impact evaluations done after projects are completed. The Development Impact Evaluation initiative based in DEC has helped, and many new impact evaluations are underway. Nonetheless, there is still much to do. The vast majority of Bank lending operations still are not properly evaluated after they are completed. Recent assessments by the Independent Evaluation Group indicate that three-quarters or more of Bank lending operations do not have impact evaluations (World Bank 2014d), although that is still an improvement compared to 15 years ago (World Bank 2012a).

The concerns go beyond the number of evaluations. The subset that is evaluated cannot be considered to be representative of the whole, as World Bank (2010) shows in the case of before-the-project evaluation, and World Bank (2012a) shows for after-the-fact evaluation. The already limited after-the-fact evaluations have been skewed in the last 10 years or so toward projects, or aspects of projects, amenable to randomized control trials (Ravallion 2009; World Bank 2012a). As a result, we see fewer evaluations of other types of projects when a simple assignment of participants and nonparticipants does not exist or when such an analysis is severely contaminated by spillover effects. There are other concerns. Evaluations tend to be biased toward short-run impacts; there have been remarkably few impact evaluations that can claim to have tested for the long-term impacts of Bank operations.

Granted, long-term evaluation can be difficult, but it is still possible: for an example, see Chen, Mu, and Ravallion (2009), an evaluation of a Bank lending operation in China. The Independent Evaluation Group has also raised doubts about how much the limited impact evaluations that have been done (inside and outside the Bank) are being used in project preparation and reporting (World Bank 2015). Moreover, the evaluations that have been done have only rarely measured impacts on poverty, even though poverty reduction is the Bank's overall goal (Goldstein 2014). It cannot be claimed that these shortcomings stem solely from technical problems or costs of doing evaluations; the problems lie elsewhere in how the Bank functions, a subject to which we will return.

From the 1990s on, it came to be understood that traditional project-specific evaluations need to be augmented by a broader assessment of public spending. There are two main reasons for taking a broader perspective. First, aid is to some degree fungible, so that aid ostensibly tied to a specific project is really just freeing up money to be spent in other areas. A rigorous cost–benefit analysis of a specific project may tell us very little about what the aid is actually funding. There is evidence of fungibility (Feyzioglu, Swaroop, and Zhu 1998), although it is less plausible in some relevant circumstances: for example, in heavily aid-dependent countries or for projects that require the external technical assistance that comes with the aid.

Second, portfolio effects arise when the multiple elements of the program package interact. The success of an education project (say) may depend crucially on whether infrastructure or public sector reform projects have worked. Evaluating each bit separately and adding up the results will not (in general) give us an unbiased estimate of the portfolio's impact (Ravallion 2016, chap. 6).

More holistic country-level approaches to assessing aid effectiveness have emerged, aiming to put each specific project in a broader public finance context. Various World Bank analytic documents (called *Poverty Assessments* and *Public Expenditure Reviews*) have played a role. However, while these analyses are useful complementary elements to cost–benefit analysis, they should not be a substitute for it. We should still know what the economic rationale is for any public project and what was learnt about its impact—both the good and the bad news.

Development Data

The World Bank has long been the one-stop shop for development data. Historically, much of this effort was through compilations of country-level data, initially in the Bank's annual *World Development Reports*, but breaking off to form the *World Development Indicators*. Since 2010, the Bank has provided open access to these data.

The Bank's data compilations are valuable, but one can also feel a degree of frustration in what has *not* been accomplished on the data production side. The sorry state of the national accounts in much of the developing world—for example, Jerven (2013) points to serious concerns about the quality of national accounts data for sub-Saharan Africa—cannot be entirely blamed on the World Bank and the IMF, but these organizations bear some responsibility. The Bank has not used its own power as much as it could to encourage governments to make public their own data,

which can help improve its quality. I recall attending a meeting with then–Bank President Robert Zoellick in 2010 at which he expressed justified alarm on realizing that the Bank provides budget support to some countries that do not provide fully public budget documents. Zoellick pushed hard on this point and budget transparency improved. But an institution committed to poverty reduction should also insist, as a prerequisite for support, that countries provide open public access to the micro- and administrative data from their own statistical offices that are needed to monitor progress against poverty.

Starting in the 1980s, the Bank's data efforts started to be more analytically driven and policy relevant. Looking at the 1979 *World Development Report*, Bank President Robert McNamara was shocked to see that only 17 developing countries had data on poverty and inequality, even though estimates of macroeconomic aggregates from national accounts were available for virtually all countries. McNamara asked his research staff to collect the missing data. With the founding of the Living Standards Measurement Study in the 1980s—which involved detailed household-level surveys of a wide range of data—as well and subsequent initiatives like the Enterprise Surveys that collect firm-level data and the Quantitative Service Delivery Surveys that collect data on health and education facilities, the Bank soon emerged as a major source of microdatasets. Bank researchers have played an important role in public microdata production.

Complementary initiatives in software development to make microdata more accessible in developing countries have greatly expanded policy and analytic capabilities; a good example for the description and analysis of microdata is the ADePT software platform (available at the World Bank website[1]). Facilitating data collection and access to relevant analytic tools should be central to the mandate of a knowledge bank.

The Bank itself also needs to be open about the data related to its own lending operations. Much data is collected in loan preparation, supervision, monitoring, and evaluation. These data can be valuable to other aid agencies and potential private financiers in facilitating learning from Bank operations—both the successes and failures. However, much of these data are not made public, and there is scope for selectivity in what is made public. Currently the incentives are weak to change this practice.

Research

Research and analytic capability is crucial to the rationale for the World Bank as a "knowledge bank." A significant share of that capability needs to be in-house, given the difficulties of structuring incentives for outsiders to deliver what is needed (Squire 2000). The Bank's research department aims to span all sectors of the Bank's work. Research is also done in some of the sectoral/regional units. Bank

[1] The ADePT software platform is available here: http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/EXTPROGRAMS/EXTADEPT/0,,menuPK:7108381~pagePK:64168176~piPK:64168140~theSitePK:7108360,00.html.

research takes many forms, ranging from project evaluations to analytic assessments of the constraints on development in specific settings.

To be a true knowledge bank, research needs adequate and secure funding. The Bank spends less on research as a share of its budget than comparable organizations, and Bank spending devoted to research has declined in real terms over recent years (World Bank 2012b). The Bank's knowledge activities (not just research) have also become more dependent on "soft money," notably trust funds.

The research function at a knowledge bank poses an organizational balancing act that needs to be more explicitly acknowledged and managed. On one side, there is a risk of "ivory-tower" researchers becoming isolated from operations. On the other side, Bank research sometimes needs to be protected by its management from the efforts of the sectoral and regional empires to influence the themes and messages of that research. A research paper that identifies deficiencies in the policies of any prominent national borrower may have a hard time getting cleared—and clearance by the Bank's country director for the country concerned is a requirement for publication. It is rare for a research paper to not be cleared, although edits are often called for. And of course, the need for this clearance is anticipated in choices made about what to research and how to present the results.

There must also be effective demand for knowledge in operations. The bulk of the Bank's senior operational staff appears to value Bank research for their work, and come to know it well (based on a survey of senior Bank operational staff discussed in Ravallion 2013). But there is a marked unevenness. The staff members working on poverty, human development, and economic policy tend to value and use Bank research more than staff in the more traditional sectors of Bank lending—agriculture and rural development, energy and mining, transport, and urban development. The latter sectors account for 45 percent of lending, but of the Bank staff who report they are highly familiar with Bank research, only 15 percent are in these sectors (Ravallion 2013). Of course, there are two sides to this problem. Demand for Bank research is interrelated with supply, and stronger incentives for learning within the Bank must come with more relevant and accessible research products.

The Bank's knowledge role should continue to include facilitating independent research outside the Bank, especially in developing countries. A good example is the Bank's support of the Global Development Network, which since its inception in 1999 supports researchers from developing countries on a competitive basis, with both financial support and by connecting researchers globally.

Policy Advice

For a knowledge bank to be credible, all parties must have confidence that the institution is not under the undue influence of powerful shareholders. At one time or another, it is likely that all of the World Bank's major shareholders and borrowers have attempted to influence Bank policies and processes. For example, some countries have been known to lobby against a Bank index of performance (such as on governance) when it ranks that country low, although such lobbying rarely appears to succeed. However, the influence of the United States has been a longstanding

concern in some quarters. The United States does have considerable power at the Bank, including in selecting the Bank's President, its weight in formal voting at the Board and in (more subtle) policy positions, and even in project implementation (for example, Kilby 2013 looks at how politics affects project preparation times). Critics question American influence on the policies advocated by the Bank in developing countries.

The bulk of this critique has focused on the Bank (and IMF) advocacy of a set of "neoliberal" economic policies that came to be known as the "Washington Consensus" (Williamson 1990). The policies included fiscal discipline, cutting generalized subsidies, tax reforms, market interest rates, liberalizing trade and foreign direct investment, privatizing state-owned enterprises, de-regulation to encourage competition, and assuring legal security for property rights. From a marketing point of view, the label "Washington Consensus" could hardly have been more damaging. The label suggests a policy agenda formed amongst an elite group in one high-income country, making the policies an easy target for some critics (for example, Broad and Cavanagh 2009).

The critics were not always well-informed about the economic rationales for those policies. There were clearly specific contexts where the policies made sense. Nor did the critics always make clear what alternative policies they had in mind and what their welfare impacts would be. For example, research has often shown that inflation is costly to poor people (for a review, see Ravallion 2016, chap. 8), so the poor have an interest in macroeconomic stability. Exaggerated claims were heard about the adverse impacts of macroeconomic adjustment on poverty; careful analysis (also considering the costs to poor people of not adjusting) often painted a more nuanced picture (for example, World Bank 1994; Jayarajah, Branson, and Sen 1996; Sahn, Dorosh, and Younger 1997).

However, some of the criticisms were valid. Early World Bank (and IMF) programs for "structural adjustment" paid too little attention to the implications for poverty reduction and human development. A welcome change in thinking within the Bank was already underway by the late 1980s. Add-on programs to "compensate the losers from adjustment" were becoming common. There was also a mounting effort to use evidence to understand the social impacts of economy-wide and sectoral policies.

The Washington Consensus was too formulaic to be credible as a policy prescription. It listed a single set of policies, but governments of developing countries could see for themselves that there were multiple paths to development success. In particular, the non-Washington Consensus route taken by China since 1980 stood out as an example for all to see. Development policy-making has become more open to what were once considered heterodox ideas, though it remains true that all policy advocates should justify their case. Theory and evidence remain no less relevant when one takes a more contextual and pragmatic approach.

An objective country-specific assessment of the binding constraints on poverty reduction should ideally guide all World Bank support. About one-quarter of total Bank lending involves what are now called Development-Policy Loans (formerly

structural adjustment loans), which are quick-disbursing loans to support a government's policy reform plans. While Development-Policy Lending operations often draw on high-level expertise within and outside the Bank, it is not clear how much influence that expertise has or how well the operations are tailored to addressing the most important constraints on development in each country. This is even less clear for the Bank's investment-lending portfolios. A series of innovations have tried to make the rationales clearer, such as in the "Country Assistance Strategy" papers. But too often, these appear to be little more than post hoc rationalizations for the lending program, rather than decisive independent analyses of what needs to be done to assure more rapid progress against poverty in the specific context. I am not the only observer to note the generally declining quality of these types of papers over time; they do not appear to be getting the attention that they once held.

Striking a balance between independent World Bank judgment and what its client countries wish to do is a continuing challenge. In 2014, the Bank introduced Systematic Country Diagnostics, in which the Bank's country teams try to identify the main development problems the country faces (and which serve as an input to the Country Partnership Framework, developed with the government). In principle, the new diagnostic tool is not confined to issues identified by the government, acknowledging the desirability of the Bank's independent view. However, the official guidelines for the new diagnostic tool say that it is to be done "in close consultation with national authorities" (World Bank 2014b, p. 1). It remains to be seen how independent the country diagnostics will be in practice, and whether politically sensitive analytics will surface in policy dialogues, especially in the large borrowing countries.

The compartmentalization of knowledge has also constrained policy advice. The Bank's sectoral silos (now called Global Practices) have not been well-suited to identifying trade-offs across sectors. More attention to trade-offs among different methods of fighting poverty is needed, and this would also be welcome for many of the Bank's clients who face hard allocative decisions.

Over the last 15 years or so, an increase in social protection spending by developing-country governments came with considerable financial support from the Bank (World Bank 2014c). This area is less attractive to the private sector (compared to infrastructure, say). But here too, the Bank's policy stances seem to strive too much for universality. Social-protection policy advocacy turned "targeting" (avoiding leakage to the "non-poor") into a fetish—oddly confusing the ends and means of social protection (Ravallion 2016, chap. 10). Lending and policy advice in this area has been dominated by a "flavor-of-the-month" approach. For a time, there was a rush to create "conditional cash transfer" schemes, providing transfer payments conditional on keeping children in school and attending to their health care. The popularity of these programs was to some degree informed by evaluations that had demonstrated impact. For example, well-documented research on the Progresa program in Mexico was very influential; on reviewing this and other evidence, a Bank research report by Fiszbein and Schady (2010) stimulated greater Bank support for conditional cash transfer schemes in numerous countries. However, conditional cash transfer advocates did not always pay proper attention to other research findings on the supply-side

delivery problems in health and education. Conditional cash transfers work less well in settings where the problem does not obviously appear to be on the demand side, given the evident failings of public service provision—failings to which Bank research has often pointed (for example, World Bank 2003).

Enthusiasm amongst practitioners ran well ahead of evaluative research for some other social policies. As one example, the weakness of local states led to well-intentioned efforts to implement Community Driven Development, in which local communities would ostensibly drive the development process rather than the state. Many development agencies, along with the Bank, provided substantial funding for community-based projects. But evaluative work soon pointed to concerns, including project capture by local elites. A more nuanced view emerged amongst researchers, which acknowledged the potential benefits of citizen participation but also warned that local states needed to be strong enough to assure that participation was effective and pro-poor (Mansuri and Rao 2013). Citizen participation is not a substitute for local state capacity. There could be a trade-off between the local-level fairness of participatory implementation and a development project's impact on poverty (Chen, Mu, and Ravallion 2009). Such trade-offs need to be taken more seriously in lending and aid, such as in poor-area development efforts.

Taking a Longer-Term Perspective on Development

World Bank policy advice needs to take a longer-term perspective on a country's development. Countries are essentially locked out of support from the development banks and most bilateral donors if their institutional environment is deemed to be too poor; in the case of the World Bank this is measured by a very low score in the Bank's Country Policy and Institutional Assessments. Once the quality of the institutional environment rises above a minimum threshold, lending and aid start to flow, with the aim (in part) of improving governance and the institutional environment more generally. This model is based on a belief that development lending and aid can improve governance (in contrast to the view of some aid critics that it promotes bad governance). External assistance eventually stabilizes when institutions are sufficiently well developed. Beyond some point, development assistance declines and eventually vanishes.

The parameters of this model are open to debate. The lack of justification for the Bank's income thresholds has been a long-standing concern—in part because the Bank's questionable criteria are widely used by other aid agencies. A more flexible approach based on relevant economic factors, such as creditworthiness and domestic capacity for redistribution, is long overdue.

But even taking the parameters as given, a feature of this model often not acknowledged properly by either aid critics or supporters is that such a model can readily yield multiple equilibria in institutional development (Ravallion 2016, Ch. 9). This has important implications for policy. For example, getting out of the low equilibrium of weak institutions—what I dub a "poor institutions trap" (PIT)—will often not be possible with only a small positive incentive for reform. As another example, fragile states could be destabilized enough to easily end up in a PIT.

This argument points to a role for the Bank in longer-term institutional development. If the World Bank were to anchor its engagement to a plan for addressing the relevant constraints in each country, its engagement would not be capricious—buffeted by short-term political shocks in its client countries or foreign-policy considerations amongst its major shareholders. To its credit, the Bank does take a longer-term perspective on development than most other aid agencies; this is evident in the attention that the Bank has given to institutional development (Birdsall and Kharas 2014) and its greater use of the recipient country's own performance management system (Knack 2013).

International Public Goods

While the World Bank is increasingly called upon to address development problems that spillover across country borders—such as pandemics and climate change—it is far from clear that it is currently well equipped for such tasks. The Bank looks for opportunities to address international public bads and has responded at times, but its present country-lending model is not well-suited to such tasks. As Birdsall (2014) points out, the Bank's $400 million Ebola response in 2014 was a fortuitous fit with the country model, rather than the systematic application of an adequately funded institutional mandate.

The Bank's new Global Practices have the capability of significant sectoral knowledge transfer across borders. The Bank also has a convening power that can help in the cross-country coordination needed in addressing global commons issues. But the required level of demand for international public goods cannot be expected to come from individual nations on their own, given the externalities involved. For the Bank to play a larger role in this area, a stronger mandate is required from its shareholders and there must be dedicated funding for global commons tasks (Morris and Gleave 2015). Birdsall (2014) suggests a new arm of the Bank is needed, or even a new institution.

It is hard to see any of this happening soon; the Bank's major shareholders have shown little enthusiasm for providing the extra capital required for new global initiatives, and many of the Bank's borrowing countries are inclined to oppose any potential diversion of funds from traditional country-based lending.

Knowledge Dissemination

There is little point in producing development knowledge that cannot be shared. A knowledge bank will naturally produce a wider range of knowledge products than a dedicated research center alone, or an academic institution. There is a role for the aforementioned "knowledge broker" function. More broadly, the task of "learning in lending" will require effort at careful documentation. Bank research should meet scholarly standards when relevant, but it should not be judged solely by narrow academic criteria. Instead, its aim must be to inform policy debates and to provide a constructively critical perspective on Bank operations. While acknowledging the differences from academic research outputs, there are a number of concerns about the Bank's current knowledge products.

First, there are quality concerns. Publication processes entail peer review, which provides a degree of quality control; the research of DEC (the chief research arm of the World Bank) tends to be published and so is subject to peer review, generally external to the Bank. However, while unpublished knowledge products customized to client needs are important to the Bank's impact on the ground, the quality of the internal review process and final output is in my experience uneven, and this should be a source of concern.

Second, the Bank's more operationally oriented knowledge products (whether published or not) have often struck me as remarkably self-referential, with rather limited signs of new knowledge entering from outside the institution. If something has not already been tried within the Bank, then it is often treated as risky—even if there is outside experience that might help evaluate that risk more clearly. Established methodologies within the Bank have a persistence that often defies innovation, new knowledge, and sometimes even old knowledge.

Third, the Bank's size and the pressures on each unit to stay big also foster knowledge products that are essentially "make-work" schemes that make little or no contribution to knowledge and so have attracted little attention. Using the very broad citation data that can be assembled from Google Scholar,[2] in Ravallion and Wagstaff (2012), my coauthor and I find that it is hard to discern more than a negligible impact for many Bank publications, though certainly not all.

Must the Lending Bank Rule?

The World Bank is not a monolithic, technocratic, poverty-minimizing agent. While eliminating absolute poverty and sharing prosperity are espoused as its overarching goals, the objectives of its staff and managers are not as well aligned with those goals as they should be. Instead, more diverse and complex motives emerge out of the Bank's governance and the multiple interests of its various stakeholders.

One important motive is to maintain and expand the institution itself. The profits from its lending have historically been an important source of revenue for Bank staffing, so it can be no surprise that the Bank's "lending culture" rewards operational staff for the volume of their lending. However, as we have seen, weak evaluative practices entail weak connectivity between Bank lending and its goal of poverty reduction. The managers/directors of the country teams have an incentive to push a high volume of lending to satisfy their bosses and ensure a decent budget for their unit, without giving sufficient consideration to the quality of that lending and how it will benefit poor people, or how it will affect the transfer of knowledge. In the process, the lending bank also generates a gauntlet of procurement rules and

[2] Google Scholar casts a broader net than other bibliographic databases, including citations by books, working papers, reports, conference proceedings, open-access journals, new, and less well-established journals. It is also more "global" in its reach, as it includes research outputs from everywhere in the world and all languages.

other administrative hurdles that absorb much staff time. Maintaining, let alone developing, the human capital of staff can be a challenge.

Concerns about the alignment of incentives in the Bank are not new. For example, this was a theme of a high-level Bank report nearly a quarter-century ago (Wapenhans 1992). Organizational changes in 1987, 1996, and 2014 sought to improve incentives for learning from lending. But with reference to the changes in 1987 and 1996, the Independent Evaluation Group concluded that: "These changes have not led to a significant change in learning from lending because they touched neither the culture nor the incentives" (World Bank 2014d, p. vii). While the new Global Practices are a promising step, all indications are that the lending culture thrives today, and still with generally weak accountability to the Bank's overall goals. Bank insiders continue to debate how to better assure that managerial choices are consistent with the Bank's overall goals (Over and Ravallion 2012; Gaarder and Bartsch 2015).

The idea of bundling knowledge with lending is still attractive to the Bank's clients. The traditional country-based model remains relevant as a means of identifying and solving pressing development problems. The complementarities with private finance point to a continuing relevance of the Bank's projects and policy support. The challenge for the Bank today is to assure that knowledge drives lending and aid, rather than simply serving them when called upon. This requires a quite fundamental change in the Bank's culture such that managerial and staff incentives are reoriented from lending to learning.

■ *For helpful comments, the author thanks Ulrich Bartsch, Tim Besley, Nancy Birdsall, Laurent Bouton, Michael Clemens, Asli Demirguc-Kunt, Shanta Devarajan, Francisco Ferreira, Marie Gaarder, Alan Gelb, Garance Genicot, Manny Jimenez, Ravi Kanbur, Steve Knack, Aart Kraay, Branko Milanovic, Rinku Murgai, Mead Over, Giovanna Prennushi, Martin Rama, Biju Rao, John Rust, Luis Serven, Lyn Squire, Jon Strand, Vinod Thomas, Dominique van de Walle, Nicolas van de Walle, Adam Wagstaff, and Bob Zoellick. The author is also grateful to the journal's managing editor, Timothy Taylor, and co-editors Enrico Morelli and Gordon Hansen for many useful comments.*

References

Birdsall, Nancy. 2014. "My Two Big Worries about the World Bank." Center for Global Development blog post, November 3. http://www.cgdev.org/blog/my-two-big-worries-about-world-bank.

Birdsall, Nancy, and Homi Kharas. 2014. *The Quality of Official Development Assistance.* Third Edition. Center for Global Development and Brookings Institution.

Broad, Robin, and John Cavanagh. 2009. *Development Redefined: How the Market Met Its Match.* Boulder: Paradigm.

Chen, Shaohua, Ren Mu, and Martin Ravallion. 2009. "Are There Lasting Impacts of Aid to Poor Areas? Evidence from Rural China." *Journal of Public Economics* 93(3–4): 512–28.

Collier, Paul, and David Dollar. 2002. "Aid Allocation and Poverty Reduction." *European Economic Review* 46(8): 1475–1500.

Deaton, Angus. 2013. *The Great Escape: Health, Wealth, and the Origins of Inequality.* Princeton University Press.

Devarajan, Shantayanan, Lyn Squire, and Sethaput Suthiwart-Narueput. 1997. "Beyond Rate of Return: Reorienting Project Appraisal." *World Bank Research Observer* 12(1): 35–46.

Englmaier, Florian, and Markus Reisinger. 2008. "Information Coordination, and the Industrialization of Countries." *CESifo Economic Studies* 54(3): 534–50.

Feyzioglu, Tarhan, Vinaya Swaroop, and Min Zhu. 1998. "A Panel Data Analysis of the Fungibility of Foreign Aid." *World Bank Economic Review* 12(1): 29–58.

Fiszbein, Ariel, and Norbert Schady. 2010. *Conditional Cash Transfers for Attacking Present and Future Poverty.* With Francisco H. G. Ferreira, Margaret Grosh, Nial Kelleher, Pedro Olinto, and Emmanuel Skoufias. Washington, DC: World Bank.

Gaarder, Marie, and Ulrich Bartsch. 2015. "Creating a Market for Outcomes: Shopping for Solutions." *Journal of Development Effectiveness* 7(3): 304–316.

Gilbert, Christopher, Andrew Powell, and David Vines. 1999. "Positioning the World Bank." *Economic Journal* 109 (459): F598–F633.

Goldstein, Markus. 2014. "Do Impact Evaluations Tell Us Anything about Reducing Poverty?" *Development Impact* blog, World Bank. http://blogs.worldbank.org/impactevaluations/do-impact-evaluations-tell-us-anything-about-reducing-poverty.

Hausmann, Ricardo, Dani Rodrik, and Andrés Velasco. 2008. "Growth Diagnostics." Chap. 15 in *The Washington Consensus Reconsidered: Towards a New Global Governance,* edited by Narcis Serra and Joseph E. Stiglitz. New York: Oxford University Press.

Jayarajah, Carl, William Branson, and Binayak Sen. 1996. *Social Dimensions of Adjustment: World Bank Experience 1980–93.* Washington, DC: Operations Evaluation Department, World Bank.

Jerven, Morten. 2013. *Poor Numbers. How We Are Misled by African Development Statistics and What to Do about It.* Ithaca, NY: Cornell University Press.

Kanbur, Ravi. 2003. "Regional Versus International Financial Institutions." In *Regional Public Goods: From Theory to Practice,* edited by A. Estevadeordal, B. Frantz, and T. R. Nguyen. Inter-American Development Bank and Asian Development Bank.

Kilby, Christopher. 2013. "The Political Economy of Project Preparation: An Empirical Analysis of World Bank Projects." *Journal of Development Economics* 105: 211–25.

Knack, Stephen. 2013. "Aid and Donor Trust in Recipient Country Systems." *Journal of Development Economics* 101: 316–29.

Kraay, Aart, and David McKenzie. 2014. "Do Poverty Traps Exist? Assessing the Evidence." *Journal of Economic Perspectives* 28(3): 127–48.

Kuziemko, Ilyana, and Eric Werker. 2006. "How Much is a Seat on the Security Council Worth? Foreign Aid and Bribery at the United Nations." *Journal of Political Economy* 114(5): 905–930.

Mansuri, Ghazala, and Vijayendra Rao. 2013. *Localizing Development: Does Participation Work?* Washington, DC: World Bank.

Morris, Scott, and Madeleine Gleave. 2015. "The World Bank at 75." CDG Policy Paper 58, Center for Global Development, Washington, DC.

Obstfeld, Maurice, and Alan M. Taylor. 2004. *Global Capital Markets: Integration, Crisis, and Growth.* Cambridge University Press.

Over, Mead, and Martin Ravallion. 2012. "Recognizing and Rewarding the Best Development Professionals." *Development Impact* blog post, October 11. World Bank, Washington, DC.

Ravallion, Martin. 2009. "Should the Randomistas Rule?" *Economists' Voice* 6(2): 1–5.

Ravallion, Martin. 2013. "Knowledgeable Bankers? The Demand for Research in World Bank Operations." *Journal of Development Effectiveness* 5(1): 1–29.

Ravallion, Martin. 2016. *The Economics of Poverty: History, Measurement, and Policy.* New York: Oxford University Press.

Ravallion, Martin, and Adam Wagstaff. 2012. "The World Bank's Publication Record." *Review of International Organizations* 7(4): 343–68.

Rosenstein-Rodan, Paul. 1943. "Problems of Industrialization of Eastern and Southeastern Europe." *Economic Journal* 53: 202–11.

Sachs, Jeffrey D. 2005. *Investing in Development: A Practical Plan to Achieve the Millennium Development Goals.* New York, NY: Millennium Project, United Nations.

Sahn, David E., Paul A. Dorosh, and Stephen D. Younger. 1997. *Structural Adjustment Reconsidered: Economic Policy and Poverty in Africa.* Cambridge University Press.

Squire, Lyn. 2000. "Why the World Bank Should Be Involved in Development Research."

Chap. 4 in *The World Bank: Policies and Structure*, edited by Christopher L. Gilbert and David Vines. Cambridge University Press.

Temple, Jonathan R. W. 2010. "Aid and Conditionality." Chap. 67 in *Handbook of Development Economics, Vol. 5*, edited by Dani Rodrik and Mark Rosenzweig. Amsterdam: North-Holland. Available at: http://www.sciencedirect.com/science/handbooks/15734471.

Wapenhans, Willi. 1992. *Effective Implementation: Key to Development Impact*. Report of the World Bank's Portfolio Management Task Force. World Bank.

Williamson, John. 1990. "What Washington Means by Policy Reform." Chap. 2 in *Latin American Adjustment: How Much Has Happened?* edited by John Williamson. Washington, DC: Institute for International Economics.

Wolfensohn, James D. 1996. "Annual Meetings Address: People and Development." October 1. http://web.worldbank.org/WBSITE/EXTERNAL/NEWS/0,,contentMDK:20025269~menuPK:34474~pagePK:34370~piPK:34424~theSitePK:4607,00.html.

World Bank. 1994. *Adjustment in Africa*. New York: Oxford University Press.

World Bank. 2003. *World Development Report 2004: Making Services Work for Poor People*. New York: Oxford University Press.

World Bank. 2005. *World Development Report: Equity and Development*. New York: Oxford University Press.

World Bank. 2010. "Cost-Benefit Analysis in World Bank Projects." Washington, DC: Independent Evaluation Group, World Bank.

World Bank. 2012a. *World Bank Group Impact Evaluations: Relevance and Effectiveness*. Washington, DC: Independent Evaluation Group, World Bank.

World Bank. 2012b. *Research at Work: Assessing the Influence of World Bank Research*. Report to World Bank's Board.

World Bank. 2014a. Annual Report 2014. Washington, DC: World Bank.

World Bank. 2014b. "Interim Guidelines for Systematic Country Diagnostic (SCD)." Report 85905. Available at: http://documents.worldbank.org/curated/en/2014/02/19227271/interim-guidelines-systematic-country-diagnostic-scd.

World Bank. 2014c. *The State of Social Safety Nets 2014*. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/2014/05/19487568/state-social-safety-nets-2014.

World Bank. 2014d. *Learning and Results in World Bank Operations: How the Bank Learns, Evaluation 1*. Independent Evaluation Group, World Bank, Washington, DC.

World Bank. 2015. *Learning and Results in World Bank Operations: Towards a New Learning Strategy, Evaluation 2*. Independent Evaluation Group, World Bank, Washington, DC.

Journal of Economic Perspectives—Volume 30, Number 1—Winter 2016—Pages 95–116

The World Trade Organization and the Future of Multilateralism

Richard Baldwin

When the General Agreement on Tariffs and Trade was signed by 23 nations in 1947, the goal was to establish a rules-based world trading system and to facilitate mutually advantageous trade liberalization. As the GATT evolved over time and morphed into the World Trade Organization in 1993, both goals have largely been achieved. The WTO presides over a rule-based trading system based on norms that are almost universally accepted and respected by its 163 members. Tariffs today are below 5 percent on most trade, and zero for a very large share of imports.

Despite its manifest success, the WTO is widely regarded as suffering from a deep malaise. The main reason is that the latest WTO negotiation, the Doha Round, has staggered between failures, flops, and false dawns since it was launched in 2001. But the Doha logjam has not inhibited tariff liberalization—far from it. During the last 15 years, most WTO members have massively lowered barriers to trade, investment, and services bilaterally, regionally, and unilaterally—indeed, everywhere except through the WTO. The massive tariff cutting that has taken place around the world, shown in Table 1, has been at least as great as in the previous successful WTO rounds. Moreover, the Doha gridlock has also not dampened nations' interest in the WTO; 20 nations, including China and Russia, have joined since 2001.

This paper begins by sketching the historical context of the original GATT agreement. It then discusses how the rules and principles behind the GATT rounds

■ *Richard Baldwin is Professor of International Economics, The Graduate Institute, Geneva, Switzerland. He is also Director of the Centre for Economic Policy Research (CEPR) in London, UK, and Editor of VoxEU.org. His email address is rbaldwin@cepr.org.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.1.95 doi=10.1257/jep.30.1.95

Table 1
Tariff Cutting Despite the Doha Deadlock

	Tariff rates in percentage points		*Change from 2001 to 2012*	
	2001	*2012*	*Percentage point difference*	*Percentage cut*
South Asia	22	13	−9	−41%
Middle East & North Africa (developing only)	19	12	−7	−38%
Sub-Saharan Africa (developing only)	14	11	−3	−19%
Latin America & Caribbean (developing only)	11	8	−4	−32%
East Asia & Pacific (developing only)	11	8	−3	−31%
World	10	7	−3	−30%
Europe & Central Asia (developing only)	8	4	−4	−49%
European Union	4	1	−2	−63%

Source: World Bank online database.
Note: Tariff rates shown are applied, simple mean, all products.

combined to create a juggernaut of political economy momentum in which nations kept joining the GATT and tariffs kept falling.

The paper then turns to the current woes of the WTO and why its magic seems to have failed in the Doha Round. Two major sets of reasons emerge in this discussion. First, the last round of GATT negotiations, the Uruguay Round, sought to generate additional momentum for free trade through broadening its focus, both in terms of more countries joining and in terms of additional areas that would be covered by the agreement. However, these steps toward broadening also required altering some of the historical rules and principles that had generated momentum toward free trade. The changes altered and may even have ended the political economy momentum of the WTO. Second, the rules and procedures of the WTO were designed for a global economy in which made-here–sold-there goods moved across national borders. But the rapid rising of offshoring from high-technology nations to low-wage nations has created a new type of international commerce. In essence, the flows of goods, services, investment, training, and know-how that used to move inside or between advanced-nation factories have now become part of international commerce. For this sort of offshoring-linked international commerce, the trade rules that matter are less about tariffs and more about protection of investments and intellectual property, along with legal and regulatory steps to assure that the two-way flows of goods, services, investment, and people will not be impeded.

It's possible to imagine a hypothetical WTO that would incorporate these rules. But in practice, the rules are being written in a series of regional and megaregional agreements like the Trans-Pacific Partnership (TPP) and Transatlantic Trade and Investment Partnership (TTIP) between the United States and the European Union. The most likely outcome for the future governance of international trade

is a two-pillar structure in which the WTO continues to govern with its 1994-era rules while the new rules for international production networks, or "global value chains," are set by a decentralized process of sometimes overlapping and inconsistent megaregional agreements.

The Historical Context for the Principles of GATT

The GATT was launched in unusual times. The demand for trade liberalization was great, because tariffs were still high from the Smoot–Hawley tariff and retaliations in the 1930s. The supply of trade liberalization was, in general terms, also great as leaders of the largest trading nations wanted to avoid the protectionist mistakes of the 1920s and 1930s. The demand for and supply of trade liberalization were also powerfully driven by the political climate in the aftermath of World War I and the outbreak of the Cold War, a setting in which world trade integration became a geostrategic issue as well as a commercial issue.

The GATT's design was heavily influenced by lessons drawn from historical trade liberalization efforts (Irwin, Mavroidis, and Sykes 2009). Pre–World War I globalization had few international organizations, supported instead by Pax Britannica. During World War II, the United States effectively became the global leader, and it wanted postwar globalization to be based on international institutions. The US Congress, however, which controls US trade policy, was refusing to bind its hands with a new international organization. Instead, trade liberalization would be buttressed by a "general agreement" but no formal organization like the International Monetary Fund. The GATT was based on several principles.

One General and Five Specific Principles

There is no definitive list of principles in the GATT and WTO, and authors differ on exactly what such a list might include (for a detailed account, see Hoekman and Kostecki 2009). However, it is useful to think of one general and five specific principles. The general principle—what might be called the constitutional principle—is that the world trade system should be rules-based, not results-based. The GATT, and now the WTO, focuses on the design, implementation, updating, and enforcement of procedures, rules, and guidelines rather than on seeking to agree upon the volume of exports or market shares. This overreaching constitutional principle is implemented with five specific principles.

1) Nondiscrimination. This rule has two aspects: nondiscrimination at the border and nondiscrimination behind the border. Nondiscrimination at the border, called "most favored nation treatment" in the WTO's circumlocutive parlance (since WTO members should treat no nation better than it treats its most favored trading partner), means that any tariff which is applied should be applied equally to all WTO members. Many exceptions are allowed (for example, free trade agreements),

but these are controlled by explicit conditions. The other aspect of nondiscrimination is called "national treatment," which is the rule that within each country, taxes and regulations should be applied evenly to domestic and imported goods.

2) Transparency. Liberalizing trade and reducing conflicts over trade is easier when the actual policies are transparent to all by having been made public.

3) Reciprocity. Nations that remove barriers to imports can expect other nations to reciprocate. Again, exceptions are made, with the most notable example being that, during the GATT era, developing nations benefited from the market opening of other nations due to the most-favored-nation provisions, but they were allowed not to reduce their own tariffs. Reciprocity also applies to retaliation. When a nation engages in a practice or policy that undoes the gain another member had from a previous agreement, the aggrieved nation has the right to reciprocate—that is, to retaliate.

4) Flexibility, or "safety valves." The founders of the GATT knew that members would occasionally be subject to irresistible domestic pressure to impose trade barriers. Rather than threatening implausibly dire consequences for such actions, the GATT allows some exceptions in which nations can at times impose trade barriers, but seeks to discipline them with various strictures and requirements for compensation.

5) Consensus decision-making. Like the other principles, this one has exceptions, but most WTO decisions are by consensus.

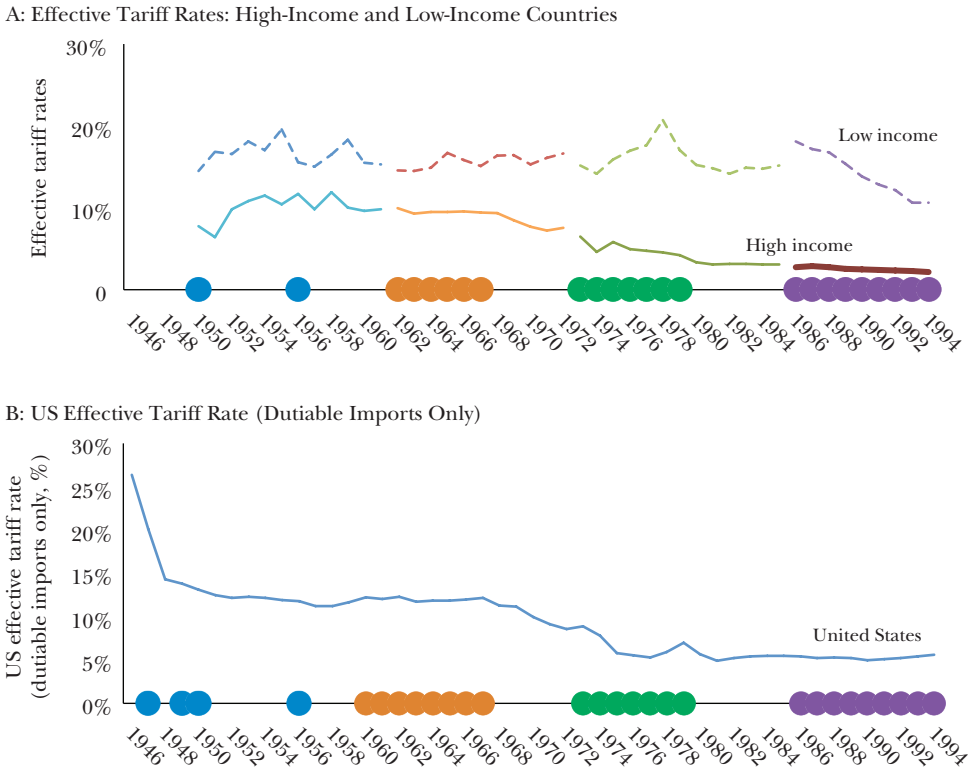
As the next section explains, interactions among these principles generated a political economy momentum that drove trade liberalization. As the following section explains, changes made in the 1990s help to explain why the momentum has ground to a halt.

A Tariff-Cutting Juggernaut

GATT is widely viewed as having facilitated the reduction of tariffs—at least in the developed nations. Systematic data on tariffs for a broad range of nations is available only from the 1980s, but a cruder measure called the "effective tariff rate"—that is, tariff revenue divided by the value of imports—has been collected back to the beginning of GATT by Clemens and Williamson (2004). An obvious problem with the effective tariff rate measure is that really high tariffs result in very low imports and so tend to get little weight in the average. In addition, the effective tariff rates for individual nations can be very noisy over periods of only a few years because they reflect both changes in tariff rates and changes in patterns of imports. Despite these well-known problems, effective rates give a reasonable general idea of tariff-cutting patterns under the GATT.

Two salient facts that emerge from Figure 1A are that low-income nations have always had higher tariffs and developed nations reduced their tariffs steadily while poor nations only started doing so in the 1980s. In looking at the figures, remember that up-and-down fluctuations in effective tariffs during periods of a few years may

Figure 1
Effective Tariff Rates, 1946–1994



Sources: Top panel, Clemens and Williamson (2004) data with author's elaboration; bottom panel, US Historical Data with author's elaborations.
Notes: In Figure 1A, the high-income nations comprise the European Union nations plus Switzerland, Norway, Japan, and Australia; the low-income nations comprise Argentina, Brazil, China, Egypt, Indonesia, India, Kenya, Korea, Mexico, Malaysia, Nigeria, Pakistan, Philippines, Thailand, and Turkey. The dots at the bottom indicate a GATT/WTO Round is in session. Figure 1B shows the effective tariff rate for the United States excluding imports with zero tariffs.

be due to shifts in the price or quantities of specific imports, along with changes in tariff policies.

Four Phases of Trade Liberalization under the GATT: 1950–1994

Figures 1A and B present averages of effective tariff rates across selected developed and low-income nations. Four phases are distinguished based on the timing of GATT Rounds.

The first phase of GATT rounds, up until 1960, began with a substantial wave of tariff cutting in the 1947 inaugural Geneva Round (visible in the US data in Figure 1B). However, the other early rounds were not focused on tariff cutting. Instead, they considered details of rules and accessions such as those of Germany

(1951) and Japan (1955). Moreover, tariffs were not the main trade hindrance to international trade in the 1950s. Instead, restrictions remaining from wartime, along with state trading and inconvertible currencies, were the binding constraints.

The second phase from 1960 up to 1972 was triggered by European regional trade liberalization (Preeg 1970). For example, the Dillon Round (1960–61) dealt with the tariff concessions European members had to make to other GATT members in compensation for the formation of their customs union.[1] The Kennedy Round (1963–1967) was, in part, an effort by the United States, Japan, and other large exporters to redress the trade diversion arising from this customs union. The decline in tariff rates in developed countries after about 1967 was in part due to GATT, but also to non-GATT steps like elimination of tariffs across much of Europe, and the US–Canada Auto Pact of 1965 which eliminated tariffs on bilateral auto trade. In this phase, regionalism and multilateralism advanced hand-in-hand. By contrast, tariffs in low-income nations did not fall since GATT rules excused them from reciprocally cutting their tariff in GATT talks. This is an exception to the nondiscrimination principle called "Special and Differential Treatment" for developing-nation members.

The third phase of trade liberalization started around 1973, and again multilateralism and regionalism advanced together. The GATT's Tokyo Round talks were launched the same year that the European Union enlarged (Britain, Ireland, and Denmark joined) and signed bilateral free trade agreements with most other West European nations. The 1970s are a period when the "effective tariff" measure can be deceiving. It looks as if substantial tariff cutting happened in developed nations although no GATT-required cuts were implemented until the round finished in 1979. The illusion arises from the 1970s price hikes that raised the import shares for oil; as developed nations had low or zero tariffs on imported oil, the relative price change looks like a cut in the average tariff. The US numbers in Figure 1B, which strip out duty-free imports (which includes petroleum in this case), show no tariff reductions at this time.

The fourth wave of multilateral and regional trade talks arose in the mid-1980s. In 1986, GATT members launched the Uruguay Round, the United States and Canada started talks about a bilateral free-trade agreement, and the European Union enlarged to include Spain and Portugal while launching its Single Market Program, which eliminated a vast range of nontariff barriers to cross-border movements of goods, services, and workers. Effective tariffs fell gently in developed nations, probably mostly due to regional rather than multilateral liberalization. For example, at the time the Uruguay Round was launched in 1986, about 40 percent of global trade took place inside free-trade areas, with about half of that within the

[1] What was later to become the European Union consisted of France, Italy, Germany, Netherlands, Belgium, and Luxembourg; the common external tariff chosen required the Benelux nations to raise tariffs they had promised not to raise in earlier talks (so-called bound rates), France and Italy to lower them, and Germany to change very little. Under GATT rules, other GATT members could demand compensation for any of the tariff rises.

European Union. The really original element in this fourth phase was the rapid tariff cutting by developing nations—but they did this outside the GATT and for reasons driven by changes in their in attitudes towards high tariffs on industrial goods (more on this change below). Developing nations also signed many regional trade agreements, like Mercosur in South America and the South African Customs Union. These had some effect on tariffs, but many developing nations lowered their multilateral tariffs at the same time as they cut tariffs with their partners in free-trade agreements (Estevadeordal, Freund, and Ornelas 2008).

As this sketch of the four phases reveals, the momentum toward cutting tariffs includes both multilateral and regional trade agreements. Thus, the underlying question is what generates this kind of political economy momentum.

The Juggernaut Dynamics of Tariff Cutting

Tariffs, like most economic policies, are the outcome of a political economy process. To explain why governments lower tariffs they previously found politically optimal to impose, the literature points to the role of trade agreements (Bagwell and Staiger 2004). The basic approach models the process as a one-time switch from a noncooperative outcome to a cooperative outcome facilitated by a trade agreement. This helps explain the initial drop in American tariffs at the start of the GATT (see Figure 1B), but a switch from one form of equilibrium to another leaves out most of the richness of how the GATT fostered multiple forms of tariff cutting over successive rounds.[2] In addition, it does not explain why developing nations acted outside the GATT to cut their tariffs starting in the mid-1980s.

More elaborate approaches to the political economy of tariff cutting draw on the intuitive, if informal, two-level game approach of Putnam (1988) as formalized by Grossman and Helpman (1995). This approach argues that governments negotiate both with special interest groups within their nation and other governments internationally. The discussion here is organized around a version of the two-level-game approach that I introduced in 1994, which I called the "juggernaut effect." It is easiest to explain in the historical context.

Before the GATT, exporters had only a very indirect interest in their nation's import tariffs. But under the GATT reciprocity principle, foreign tariff levels became linked to domestic tariff levels. Of course, this connection only held for developed nations who followed the reciprocity principle. In a way, the GATT's success was not due to the international deal itself. It was due to the way the principles behind the international deal altered domestic political realities in developed-nation members. Also, remember that developing-nation governments were excused from reciprocity by the Special and Differential Treatment rule, and thus faced the same array of domestic pro- and anti-tariff special interests before, during, and after each GATT

[2] Lockwood and Zissimos (2004) propose a Bagwell–Staiger-type model where tariff cutting is gradual due to retaliation limits that are, in their model, essential for supporting reciprocal tariff cutting.

round. In theory and practice, this meant that they did not lower tariffs that they previously found optimal to impose.

In the juggernaut story, the first round of tariff cuts creates political economy momentum. As tariffs drop, pro-tariff import-competing firms face additional international competition. Many of them shrink, become less profitable, and even go out of business. Conversely, foreign tariff cutting boosts exporters. They expand and become more profitable. In this way, a one-off tariff cut weakens protectionist forces and strengthens liberalization forces from a political economy perspective. A few years down the road, when another multilateral GATT round is launched, the altered political economy power of importers and exporters comes into play. As before, exporters have an incentive to fight for domestic tariff cuts due to the reciprocity principle, and import-competing firms have an incentive to fight against them. But since the anti-liberalization camp is systematically weaker and the pro-liberalization camp is systematically stronger than during the last round, all the governments playing reciprocally find it politically optimal to cut tariffs again. As these fresh tariff cuts are phased in, the exit of import-competing firms and entry of exporters again reshapes the political landscape inside each participating nation, and the cycle restarts. The juggernaut rolls forward.

This dynamic also suggests an explanation for why multilateral and regional tariff-cutting progressed in tandem. Once the original tariff cuts weaken protectionists and strengthen liberalizers, governments find it optimal to lower tariffs both multilaterally and regionally.

A related political economy dynamic is that regional trade agreements can kick-start multilateral trade liberalization. For example, a quick eyeballing of Figure 1B, the US effective tariff rates, suggests that the juggernaut had run low on momentum by the end of the 1950s. However, when the countries of Western Europe began cutting their intra-European tariffs from 1959, the resulting tariff discrimination aroused the concerns of exporters in the United States, Japan, and Canada.[3] At that time, North America and Japan both sent roughly one-third of their exports to Europe, and their firms feared losing these markets to European firms who enjoyed zero tariffs inside the customs union. As the impact of the discrimination would be reduced by lower EU most-favored-nation tariffs, North American and Japanese exporters lobbied for a GATT Round as a way of countermanding the discrimination. A similar thing may have happened when the European customs union was enlarged in 1973—the same year that the Tokyo Round talks started.

Avoiding Backsliding: Binding Plus Allowing Retaliation

The GATT had other mechanisms to keep this gradual, mutually advantageous tariff cutting on track. After all, the "juggernaut" process of political economy

[3] The GATT–European Economic Community link is explicit in President John F. Kennedy's "Special Message to the Congress on Foreign Trade Policy" (January 25, 1962, http://www.presidency.ucsb.edu/ws/?pid=8688).

momentum can work in reverse—as it had in the 1930s. Thus, the GATT process included a set of rules designed to make political reversals difficult for individual members. One rule was the principle that a nation's past tariff cuts were "bound" in the sense that previously agreed tariff levels were not open to further negotiation. Moreover, a nation's partners could retaliate against any violation of such tariff "bindings" by raising their own tariffs against the violating nation's exports. The effect was to ensure that each nation's exporters would be punished for any backsliding by its own government. This gave exporters an incentive to push their government to respect the bindings. Notice that this design element did not depend on the nation's own government; instead, it was enforced by the risk that foreign governments would retaliate by raising tariffs.

Three Escape Hatches

How could the many countries of the GATT reach agreements while working on a consensus principle? One answer is that some escape hatches were historically allowed, which made it easier for members to agree to the tariff cuts in the first place.

As one example of an escape hatch, a variety of GATT practices on "Special and Differential Treatment" meant that developing nations were not subject to GATT disciplines. They were exempted from an expectation of reciprocally cutting their tariffs, and they could mostly ignore any GATT rules with which they didn't agree. In short, the low-income nations that were part of GATT could typically follow a policy of "don't obey, don't object." However, being excused from reciprocity did not mean the developing nations were indifferent to the GATT's success. The GATT's most-favored-nation principle meant that the tariff cuts agreed among the developed nations were automatically extended to developing-nation exporters. They were free riders who liked the ride. The developed countries were mostly happy to allow this free riding because developing-nation markets were, at the time, rather insignificant.

A second kind of escape hatch emerged in the 1960s and 1970s during the Tokyo Rounds, in which negotiations on trade rules were undertaken by the so-called "codes" approach. In this approach, each set of rules agreed upon was adopted in the form of a code that would be binding only for those members that voluntarily signed them—which in practice typically meant the developed nations. For example, during the Tokyo Round a number of issues beyond tariffs (such as restraints on production subsidies) were put on the agenda using the "codes" approach; many of these issues involved new forms of protection that had arisen in the 1960s and 1970s to offset competitive effects of earlier tariff cuts (Baldwin 2009, 1970). However, the principle of nondiscrimination meant that countries that did agree to these codes were (mostly) obliged to extend the rules to all GATT members, even those that did not sign the codes.

A third escape hatch arose because the GATT dispute settlement system wasn't strong enough to enforce compliance. Disputes were brought before a panel whose rulings were reviewed by a group of members that included the disputing parties.

According to the consensus principle, the Panel ruling was only accepted if all parties agreed. For example, in 1959 the European Free Trade Association (EFTA) nations wanted free trade among themselves, but only on industrial goods. In 1965, the United States and Canada wanted to liberalize bilateral trade in the auto sector. When GATT panels were formed to investigate the "GATT-legality" of these limited free trade agreements, the EFTA nations and the United States blocked the panel from reaching a conclusion.

Of course, if GATT members had extremely diverse preferences, escape hatches like blocking the dispute resolution process could have become a main exit, thus rendering the rules useless. But instead, the combination of a dispute procedure with an escape hatch facilitated agreements by allowing GATT members to be satisfied with wording that could be described as "constructively ambiguous." The GATT's quasi-legal dispute mechanism with escape hatch could be relied upon to settle disputes, or at least to help frame future negotiations aimed at clarifying ambiguities if and when such clarification proved important.

Causality

The story as told hereto has been of the GATT causing tariff cuts, but how do we know that it was not a third effect causing both GATT membership and tariff cutting? The prima facie evidence is clear, even if the econometrics has not been done due to the lack of high-quality historical tariff data. Two types of tariffs were not subject to the juggernaut "treatment"—all developing-nation tariffs, and agricultural tariffs of all GATT members. Neither set of tariffs fell during the GATT days: agricultural tariffs because they were not on the negotiating table, and developing-nation tariffs because they were excused from reciprocal cuts. This suggests that no third factor was causing tariff-cutting pressures across the board; instead, the juggernaut treatment only worked on the tariffs to which it was applied.

Refueling the Juggernaut, But Closing the Escape Hatches

By the 1970s, tariffs in the developed nations were already fairly low—at least on the products on which they had been willing to negotiate. Agriculture and labor-intensive industrial goods like clothing had been explicitly taken off the bargaining table when the agendas were set for the earlier GATT rounds. In this way, the GATT liberalization resulted mostly in tariff cutting in areas that were of most interest to developed nation exporters, basically industrial goods. Developing-nation exporters, who didn't have any "skin in the game" due to Special and Differential Treatment, were often disappointed in the lack of liberalization of agriculture and labor-intensive manufactures.

To refuel the trade liberalization juggernaut, the developed nations that had mostly driven the GATT process decided to broaden the agenda. The process started during the Tokyo Round with the "codes" approach to including nontariff issues in the negotiations. Then with the Uruguay Round starting in 1986, new areas of interest to exporters in developed nations were put on the negotiating table, notably intellectual property issues, restrictions on foreign investment, and exported services issues. These areas came to be known as TRIPs (Trade-Related

Intellectual Property), TRIMs (Trade-Related Investment Measures), and services, respectively. Additionally, two sectors still marked by high tariffs—agriculture and clothing—were put on the table to fuel the interest of agriculture exporters and low-wage exporters. It was hoped this constellation of new issues would refuel the juggernaut by rebalancing interests along North and South lines. Northern exporters were to gain from new rules and new market access in TRIPs, TRIMs, and services, while Southern exporters were to gain from freer trade in food and clothing. However, the dynamics of the negotiations and the increasing importance of emerging market economies meant that as the agenda was broadened, some of the earlier escape hatches were closed up.

For example, industrial nations' domestic laws already assured intellectual property protection for foreigners, so the expected gains for intellectual-property exporters from developed countries would come primarily from getting developing nations to adopt the standards of developed countries on patents, copyrights, and the like. During the Uruguay Round, developed countries feared that their opening of agriculture and textile markets would be pocketed by developing nations, while new disciplines on TRIPs, TRIMs, and services would be picked apart. A voluntary codes approach just would not do for a deal balanced in this way. The developing nations most likely to be affected would be those most likely to opt out. As a result, the Uruguay Round ended up including a feature called the Single Undertaking. All members, developed, and developing alike—even those that had not participated actively in the negotiations—were obliged to accept all the Uruguay Round agreements as one package. The basic outlines of the package-deal approach had been discussed in December 1991 (Croome 1995). Nevertheless, it clearly came as a surprise to many developing country members, especially those that did not follow the Uruguay Round through its eight years of twists and turns.

In addition, because the new areas involved considerable ambiguity and newness, members participating in the Uruguay Round negotiations decided it was necessary to greatly reduce the wiggle room in the dispute procedure. Both North and South feared that exporters' gains in the new areas might be offset by murky forms of protection or slippery national interpretations of the rules. For example, many governments in emerging economies were concerned that the United States was prone to taking unilateral action against whatever the US government considered to be an unfair trade practice under Section 301 of the Trade Act of 1974 (Keohane and Nye 2001). The Uruguay agreement eliminated the possibility of blocking the initiation of a dispute resolution or adoption of a panel ruling, and applied this to all the areas in the Single Undertaking. The new adjudication procedure welded shut the earlier escape hatch.

Win-Win Multilateral Cooperation

From its start in 1946 until it was superseded by the World Trade Organization in 1995, the GATT promoted win-win multilateral cooperation by setting up what Douglass North would call an "institution"—constraints that guide political and economic interactions consisting of formal rules and informal restraints. The

principles of the GATT fostered a self-enforcing pattern of cooperation and success. As the GATT's liberalization process started working its magic, exports of manufactured goods boomed—growing twice as fast as the production of manufactured goods from the late 1960s until just before the collapse of trade in 2009. Booming trade and incomes strengthened the belief of GATT members that following the code of conduct was good policy. As nations and interest groups came to expect that the rules would be respected, they adopted behaviors that conformed to the rules, thus making compliance almost automatic.

The Woes of the WTO

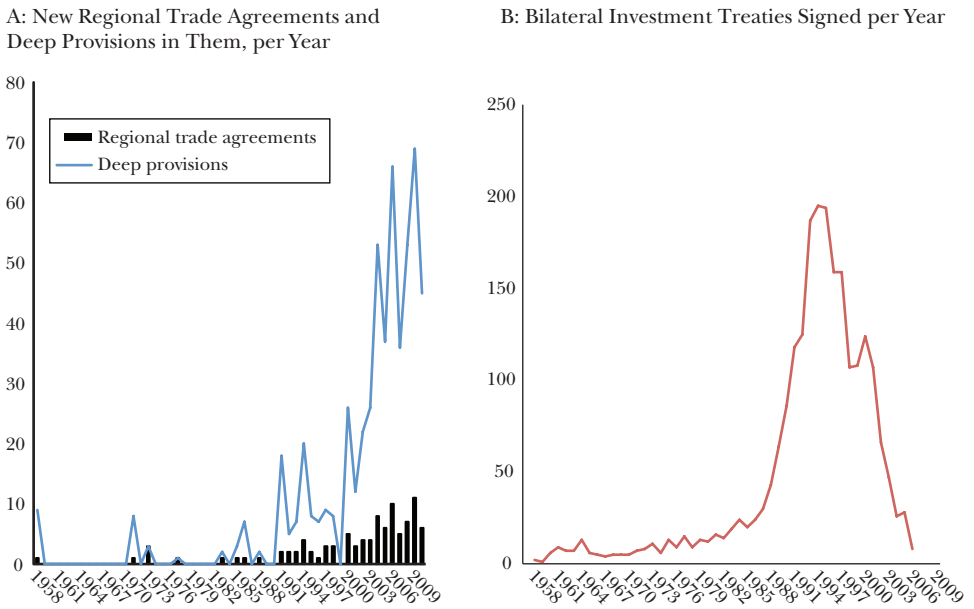
A performance review of the WTO would produce an unbalanced report card. Little progress has been made on the trade liberalization front for almost two decades, since a handful of agreements in 1997. The Doha Round that started in 2001 is stalled. Of the WTO's functions, only the dispute settlement mechanism would receive a high performance score. Why did the GATT trade liberalization magic stop working for the WTO? I consider both external and internal reasons, and then consider the implications for multilateral and regional trade talks.

External Sources

The most commonly cited causes of the WTO's difficulties involve the lost dominance of the advanced economies. This occurred in two ways. First, as discussed above, the GATT was all about exchanges of market access, so market-size was the coinage of the realm. In the GATT period, the United States, European Union, Japan, and Canada—known as the Quad—dominated on this metric, accounting for two-thirds of world imports. The rapid growth of emerging economies changed this. Today, the Quad accounts for only half of world imports. Second, the sheer number of developing country members has shifted power in the organization and made talks more difficult. Since the last successful GATT/WTO negotiation was launched in 1986, over 70 developing nations have joined, about half of them since the WTO was created. Importantly, this includes China who rejoined in 2001 (having quit two years after joining in 1948).

In theory, more member nations does not necessarily hinder tariff cutting: after all, more nations could mean more demand and more supply for better market access. In practice, however, developing countries became active in more new defensive coalitions (that is, groups interested in preventing better access to their own markets) than in new offensive coalitions (groups interested in getting better access to foreign markets) (Patel 2007). The reason is straightforward. The reciprocity principle and small size of most developing markets limited their ability to ask foreigners to open up their markets. Hence, such countries had little to gain from new offensive coalitions. The consensus principle, by contrast, gave developing-nation coalitions a good deal of blocking power, which they used to block efforts to open their most politically sensitive markets.

Figure 2
Number of Regional Trade Agreements, Deep Provisions, and Bilateral Investment Treaties



Sources: WTO RTA database (left) and UNCTAD online data (right).
Notes: Deep provisions are defined as beyond tariff cutting; see Baldwin (2012) for details. The provisions counted as deep include those that constrict nation laws on foreign investment, intellectual property rights, regulatory convergence, short-term movement of managers and technicians, and capital flows.

Regionalism also created challenges. Regional trade agreements have always been part of the trade governance landscape. From around 1990, however, they played a very different role as the number of agreements skyrocketed. As all of these involved tariff cutting that would otherwise have had to be funneled through the WTO, and as all of these took up political economic "capital," the rise of regionalism probably made it harder to conclude the Uruguay Round. Concluding the Doha Round would probably be easier if, when it comes to trade liberalization, the WTO was the "only game in town."

Many of these new regional trade agreements were "deep" in the sense of Lawrence (1996), meaning they went beyond tariff-cutting and included legally binding assurances aimed at making signatories more business-friendly to trade and investment flows from other signatories (recall that the GATT agreements are not legally binding). This can be seen in WTO data that codifies the content of regional trade agreements (based on seminal work by Horn et al. 2010). Figure 2A (left panel) shows the flow of new agreements and the flow of new deep provisions in them (according to my classification that picks out provisions related to offshoring). At about the same time, an old form of economic integration agreement became

very popular, the bilateral investment treaty (Figure 2B, right panel). Basically, these are concessions of sovereignty undertaken to encourage inward investment. For example, signatories usually commit to resolve investor–state disputes in a forum based in Washington, DC, rather than in national courts. In their heyday, scores of bilateral investment treaties were signed annually. By the late 1990s, most developing nations had already signed them with their major investment partners, so the number fell off sharply. There are now over 3,000 such agreements in existence.

The boom in the investment treaties and deep provisions did not create a direct competitor to the WTO. But they provide revealed-preference evidence that many WTO members were looking for disciplines that went far beyond the "shallow" disciplines included in WTO talks. In other words, the demand for policy reforms shifted away from the sort of disciplines that the WTO was set up to negotiate.

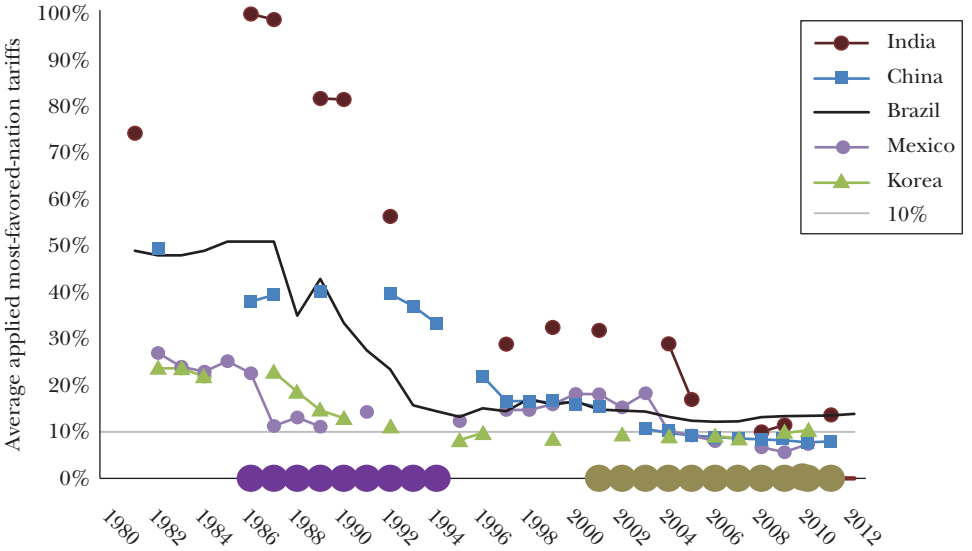
A different challenge came from unilateral tariff-cutting by developing nations. The rise of offshoring opened a new pathway to industrialization. The old, import-substitution path meant building a supply chain at home in order to become competitive abroad. High tariffs were often viewed as part of this process. The new offshoring-led path involved joining an international production network to become competitive, and then industrializing by expanding the quantity and range of tasks performed (Baldwin 2011). In this new development model, tariffs hinder rather than help industrialization, so developing-nation tariffs started to fall rapidly independently of WTO talks, as shown in Figure 3. To maintain flexibility, the developed nations did not "bind" the tariffs in the WTO even when they lowered them on a nondiscriminatory basis.

Because two-way tariff cutting had been the main fuel for the political economy juggernaut of trade liberalization, this unilateralism made multilateral talks less attractive to many developed members whose exporters saw their sales to developed nations boom even as Doha Round staggered from failure to failure. Why fight domestic protectionists at home when foreigners were lowering their tariffs unilaterally?

Internal Sources

These external challenges were magnified by big changes in the way WTO talks were organized, as opposed to those under the GATT. To put it bluntly, GATT multilateral negotiations involved the Quad (the United States, European Union, Japan, and Canada) bargaining among themselves over tariff cuts that they allowed the developing-nation members to free ride upon. WTO negotiations, by contrast, require binding tariff cuts and other policy commitments from all but the poorest members. In a political economy sense, the WTO and GATT are quite different international organizations. Specifically, as the Doha Round results would be binding equally on every member unless explicit exceptions were made, the "don't obey don't object" option that developing nations had under the GATT was cancelled under the WTO (Oyejide 2002). Not surprisingly, they have been far more vocal in the Doha Round than they were in GATT rounds, objecting to provisions that threatened their domestic interests.

Figure 3
Unilateral Tariff Cutting by Emerging Markets from the Mid-1980s



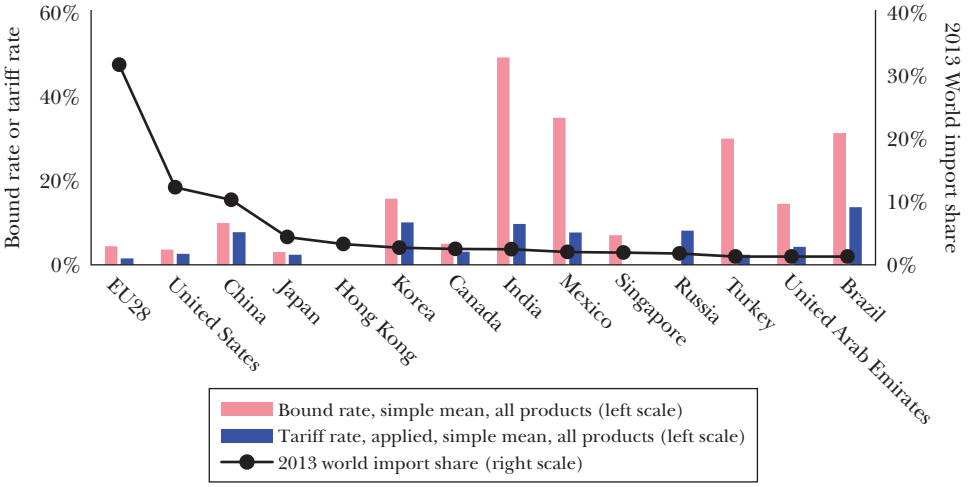
Source: Author's elaboration of World Bank online data.
Notes: Simple average of applied most-favored-nation tariffs, all products. The dots at the bottom indicate a GATT/WTO Round is in session.

Implications for Multilateral Trade Talks

The impact of these challenges was not immediately apparent. In the years following the 1994 agreement that set up the WTO, multilateral talks worked much as before. A few bits of leftover business, like the 1997 Financial Services Agreement and Information Technology Agreement, were handled in the usual fashion in negotiations led by developed economies. But as the Doha Round got underway, the world discovered that the GATT's juggernaut magic would not work in the WTO. Specifically, the external and internal challenges had three momentous implications for the WTO multilateral negotiations.

First, multilateral negotiations under the WTO are more difficult. As explained above, the Single Undertaking principle meant that instead of four veto-players (the Quad nations of the United States, European Union, Japan, and Canada) and dozens of free riders as under the GATT, the Doha Round has more than 100 potential veto-players. Second, business interest in the Doha Round is much less forceful. The agenda for the talks, set in 2001, focused on tariff cutting in industrial goods and trade distortions in agricultural and service sectors. Industrial trade accounts for 80 percent of all trade, but business interest was dampened by the fact that tariffs in the Quad nations were already low, and those in the major developing nations had been lowered unilaterally. From a WTO perspective, the exporters of developed nations were now the free riders on unilateral tariff cutting by emerging

Figure 4
Bound and Applied Rates for 14 Largest Importers, 2013



Source: World Databank for tariffs; WTO online database for imports, with author's elaboration.
Note: The nations listed above accounted for 80 percent of world imports in 2013.

markets. This greatly reduced their interest in lobbying their own governments for a Doha deal.

Second, a particular detail of WTO procedures has made unilateral tariff-cutting a major problem for the Doha Round. Following long-standing practice, WTO tariff-cutting talks focus on "bound" tariff rates, not applied rates. For many WTO members, actual applied rates are so much lower than the bound rates that the proposed Doha cuts would only reduce the distance between bound and applied tariffs, without actually lowering the applied rates. Figure 4 illustrates the issue. Bound and applied rates are shown for the 14 largest importers who together account for 80 percent of world imports. These are the markets to which exporters want access. Five decades of GATT talks had already lowered bound rates in the developed economies to less than 3 percent on average. In most of the large developing nations, bound rates are quite high, but applied rates are lower. Even in China, the third-largest global market for exports, the applied rate is about 8 percent. If the developing nations had not lowered their applied rates so much below their bound rates, developed nation exporters would have had something to fight for.

Similarly in agriculture, the biggest protectionists—the European Union and Japan—unilaterally lowered distortions for purely domestic reasons. The political power of rich-nation farm lobbies has dropped as farm populations have fallen and awareness has risen about the fact that most farm support goes to wealthy landowners and agri-corporations. The European Union broadly switched its agriculture support policies to non-trade-distorting forms and basically eliminated export

subsidies in major reforms that took place in 2003, 2008, and 2013. Japan still has astronomical tariffs on a handful of products like rice, but it too is shifting unilaterally towards non-trade-distorting policies with major reforms in 2003 and 2007. While agriculture trade is hardly free and fair (and the United States increased trade distortions with its 2014 US Farm Bill), the mercantilist gain from a conclusion of the Doha Round is clearly lower in 2016 than it would have been in 2001. Moreover, a number of emerging markets have deployed some of their newfound wealth in the form of new trade-distorting agriculture policies of their own. They are, in essence, reacting to exactly the same rural–urban domestic politics that produced agriculture protection in the United States, European Union, and Japan. This has created new opponents to agricultural trade opening.

Third, the rise of offshoring has created a political economy demand and supply for disciplines that underpin international production networks. As these disciplines were not included in the Doha Round's 2001 agenda, and dozens of WTO member have vetoed all moves to expand the agenda, the supply and demand are meeting outside the WTO—mainly in the deep regional trade agreements and bilateral investment treaties. The rapid rise in production unbundling—sometimes called "global value chains"—has meant that the world's most dynamic trade involves a nexus of trade in goods, services, know-how, physical investment, key personnel, and financial capital. Many developing nations sought and are still seeking to attract this offshoring activity. Firms in the high-income nations are interested in providing it—as long as they have assurances that host nations will respect their tangible and intangible property rights, and ensure that the necessary flows of goods, services, investment, capital, and people will be unimpeded.

These assurances have been provided in dozens of deep bilateral and regional trade agreements and in the bilateral investment treaties. This "spaghetti bowl" (as it is sometimes called) of intertwined agreements is clearly not optimal for international production sharing. As a result, a number of so-called megaregionals like the Trans-Pacific Partnership and Trans-Atlantic Trade and Investment Partnership have emerged to multilateralize some of the disciplines at a regional level. To understand this process, Ethier (1998) presents a model of how development led by foreign direct investment could affect multilateralism. In short, the political economy switched from "my market for yours" to "my factories for your reform"—that is, developing-nation trade liberalization and pro-business reforms in exchange for production facilities from developed nations.

The Future of Multilateralism, Regionalism, and the WTO

The WTO is a pillar of multilateral economic governance, as was the GATT before it. Its prime mission is to establish rules of the road and facilitate negotiation of mutually advantageous trade liberalization. In the main, the WTO can claim "mission accomplished." It oversees a set of near-universal norms for rule-based trade, and it runs a dispute settlement mechanism that routinely

arbitrates disputes and issues rulings that are universally followed even though it has no direct enforcement power. Most telling of all, nations vote with their feet by joining the WTO, even though the requisite reforms typically involve high domestic political costs.

However, the WTO seems frozen in time. The last updating of its rulebook and its last major trade liberalization came in 1994, when Bill Clinton, Gerhard Schroeder, Hashimoto Ryutaro, and Li Peng were in power and the Internet barely existed. The current WTO talks, the Doha Round, are focused entirely on 20th-century issues such as tariffs on industrial and agricultural goods, along with trade-distorting policies in agriculture and services.

While a couple of small agreements have been completed, the Doha Round is in its 15th year and nowhere near done. This 15-year fail trail, however, has not stalled global trade opening and rule-writing. For 20 years, new rule-writing and trade liberalization has proceeded apace along three axes—all of them outside the ambit of the WTO. First, a great deal of tariff cutting has been done unilaterally by WTO members, especially developing-nation members. Second, new disciplines on international investment—flows that are now intimately entwined with trade in goods and services—have been established by a network of over 3,000 bilateral investment treaties. Third, the new rules and deep disciplines that have underpinned the rapid expansion of offshoring and the internationalization of production have been written into deep regional trade agreements, especially those between advanced and emerging economies.

These observations invite two questions: Is the lack of multilaterialism worrisome? What is the future of the WTO and multilateralism?

Is the Lack of Multilateralism Worrisome?

Two decades ago, the explosion of bilateral deals shown in Figure 2B sparked a debate on multilateralism versus regionalism. Authors such as Bhagwati (1993a, b) decried regionalism as dangerous. He pointed to a "small-think" danger—that the inefficiencies of trade diversion would diminish welfare—and a "big-think" danger—that regionalism would block the path to global free trade.

As it turned out, global tariff-cutting since the rise of regionalism has proceeded as quickly as ever, but outside the WTO (as shown earlier in Figure 1). As a result, the specter that regional trading agreements would inefficiently divert trade never really appeared. Measures based on detailed tariff data show that little of world trade is affected by tariff preference margins of 5 percent or more (WTO Secretariat 2011). After all, the most-favored-nation tariffs are zero or very low on most of the world's large trade flows, and so bilateral and regional trade agreements provide a relatively small incentive to divert trade. Where tariffs remain high, bilateral and regional trade deals tend to exclude such "sensitive" items, so no preference is created either. Overall, the econometric evidence suggests that trade diversion due to bilateral and regional agreements is not a first-order concern in the world economy (Estevadeordal, Freund, and Ornelas 2008; Acharya, Crawford, Maliszewska, and Renard 2011).

As for the systemic, big-think danger, it is hard to know what would have happened if somehow nations had not signed the hundreds of bilateral agreements that they did. But one thing is clear. The rise of preferential tariffs within bilateral and regional agreements has not blocked the path to overall global tariff-cutting. Virtually all of the developing-nation WTO members who engaged in bilateral, discriminatory liberalization have simultaneously been engaged in unilateral, nondiscriminatory liberalization.

Importantly, the trade creation/diversion concern only applies to bilateral and multilateral liberalization that is truly discriminatory against trade from countries not included in the agreement. However, many of the deep regional trade agreement provisions concern matters where discrimination is impractical. Such disciplines impinge upon corporations, services, capital, and intellectual property, and in these areas it is difficult to write rules that identify the nationality of such things in a way that clever lawyers cannot get around. For example, the Japan–Thailand regional trade agreement allows Japanese banks to sell certain financial services in Thailand. But since it is difficult to determine which banks are Japanese, the agreement grants the privilege to any bank registered and regulated in Japan—which makes most large US and EU banks "Japanese" for the purposes of the agreement. This phenomenon of "soft preferences" also arose from the EU's Single Market program (which is the biggest and deepest of all regional trade agreements). As it turned out, many EU reforms were helpful to non-EU firms even though their nations were not signatories.

Future of Multilateralism

The WTO's paralysis in the face of frenetic tariff cutting and rule writing outside the organization can be attributed to two factors. First, the Doha Agenda was set for a world economy that is no longer with us. If Doha had been concluded in a few years as planned, the juggernaut effect might have worked. But with the rise of China, the rise of offshoring, and the rise of unilateralism, the negotiating items on the Doha agenda no longer provide a win–win bargain for all. Second, the natural step of expanding the WTO agenda to include some of the disciplines routinely agreed in deep regional trade agreements is blocked by nations who have been largely left aside by the rise of offshoring. They feel that they were promised, in 2001, a "rebalancing" that would involve reduced barriers to exports of agricultural and labor-intensive goods. Until they get their rebalancing, they have been willing to veto an expansion of the agenda.

Since important network externalities can be won by moving away from bilateralism and towards multilateralism when it comes to some deep provisions that are commonly found in regional trade agreements, the WTO's paralysis has led to plurilateral deals being done elsewhere. The thousands of bilateral investment treaties, for instance, are not all that different, and so network externalities could be realized by melding them together. The emergence of so-called megaregionals like the Trans-Pacific Partnership and Trans-Atlantic Trade and Investment Partnership should be thought of as partial multilateralization of existing deep disciplines

by sub-groups of WTO members who are deeply involved in offshoring and global value chains.

The megaregionals like the Trans-Pacific Partnership and Trans-Atlantic Trade and Investment Partnership, however, are not a good substitute for multilateralization inside the WTO. They will create an international trading system marked by fragmentation (because they are not harmonized among themselves) and exclusion (because emerging trade giants like China and India are not members now and may never be). Whatever the conceptual merits of moving the megaregionals into the WTO, I have argued elsewhere that the actual WTO does not seem well-suited to the task. First, as mentioned, the WTO seems incapable of getting beyond the Doha Round and incapable of addressing deep disciplines until it does. Second, a situation where China, India, and other large emerging markets stay outside the megaregionals may prove to be stable. The "soft preferences" arising from the megaregionals may not prove very damaging to large outsiders who can use their market size and unilateral harmonization to offset the negative effects. For example, those European outsiders who decided to stay out of the EU could still make adjustments and live with the soft preferences. A domino effect, however, is likely to draw in smaller outsiders wishing to participate in the international production networks inside the megaregionals.

What all this suggests is that world trade governance is heading towards a two-pillar system. The first pillar, the WTO, continues to govern traditional trade as it has done since it was founded in 1995. The second pillar is a system where disciplines on trade in intermediate goods and services, investment and intellectual property protection, capital flows, and the movement of key personnel are multilateralised in megaregionals. China and certain other large emerging markets may have enough economic clout to counter their exclusion from the current megaregionals. Live and let live within this two-pillar system is a very likely outcome.

■ *This paper draws on a several of my unpublished policy essays that have been posted in the CEPR policy discussion papers, Policy Insights (Baldwin 2010, 2011, 2012), and an unpublished paper I wrote for the OECD, Baldwin (2014).*

References

Acharya, Rohini, Jo-Ann Crawford, Maryla Maliszewska, and Christelle Renard. 2011. "Landscape." In *Preferential Trade Agreement Policies for Development: A Handbook*, edited by Jean-Pierre Chauffour and Jean-Christophe Maur, 37–76. Washington, DC: World Bank.

Bagwell, Kyle, and Robert W. Staiger. 2004. *The Economics of the World Trading System*. MIT Press.

Baldwin, Richard. 2010. "Understanding the GATT's Wins and the WTO's Woes." CEPR Policy Insight No. 49.

Baldwin, Richard. 2011. "Trade and Industrialisation after Globalisation's 2nd Unbundling: How Building and Joining a Supply Chain are Different and Why it Matters." NBER Working Paper 17716.

Baldwin, Richard. 2012. "WTO 2.0: Global Governance of Supply-Chain Trade." CEPR Policy Insight no. 64, December, Center for Economic Policy Research.

Baldwin, Richard. 2014. "Multilateralising 21st Century Regionalism." Paper prepared for the OECD conference "Global Forum on Trade Reconciling Regionalism and Multilateralism in a Post-Bali World." http://www.oecd.org/tad/events/OECD-gft-2014-multilateralising-21st-century-regionalism-baldwin-paper.pdf.

Baldwin, Robert E. 1970. *Nontariff Distortions of International Trade*. Washington, DC: Brookings Institution.

Baldwin, Robert E. 2009. "Trade Negotiations within the GATT/WTO Framework: A Survey of Successes and Failures." *Journal of Policy Modeling* 31(4): 515–25.

Bhagwati, Jagdish N. 1993a. "The Diminished Giant Syndrome." *Foreign Affairs*, Spring, pp. 22–26.

Bhagwati, Jagdish. 1993b. "Regionalism and Multilateralism: An Overview." In *New Dimensions in Regional Integration,* edited by Jaime de Melo and Arvind Panagariya, 22–51. Cambridge University Press.

Clemens, Michael A., and Jeffrey G. Williamson. 2004. "Why Did the Tariff–Growth Correlation Change after 1950?" *Journal of Economic Growth* 9(1): 5–46.

Croome, John. 1995. *Reshaping the World Trading System: A History of the Uruguay Round*. World Trade Organization.

Estevadeordal, Antoni, Caroline Freund, and Emanuel Ornelas. 2008. "Does Regionalism Affect Trade Liberalization toward Nonmembers?" *Quarterly Journal of Economics* 123(4): 1531–75.

Ethier, Wilfred. 1998. "Regionalism in a Multilateral World." *Journal of Political Economy* 106(6): 1214–45.

Grossman, Gene, and Elhanan Helpman. 1995. "Trade Wars and Trade Talks." *Journal of Political Economy* 103(4): 675–708.

Hoekman, Bernhard M., and Michael K. Kostecki. 2009. "The Political Economy of the World Trading System: The WTO and Beyond," 3rd edition. Oxford University Press.

Irwin, Douglas, Petros Mavoridis, and Alan Sykes. 2009. *The Genesis of the GATT*. Cambridge University Press.

Keohane, Robert O., and Joseph S. Nye. 2001. *Power and Interdependence*, 3rd edition. New York: Addison-Wesley Longman.

Lawrence, Robert Z. 1996. *Regionalism, Multilateralism, and Deeper Integration*. Washington, DC: Brookings Institution.

Lockwood, Ben, and Ben Zissimos. 2004. "The GATT and Gradualism." Econometric Society 2004 North American Summer Meetings 607.

Oyejide, T. Ademola. 2002. "Special and Differential Treatment." Chap. 49 in *Development, Trade and the WTO: A Handbook* edited by Bernard Hoekman, Aaditya Mattoo, and Philip English. Washington, DC: World Bank.

Patel, Mayur. 2007. "New Faces in the Green Room: Developing Country Coalitions and Decision-Making in the WTO." GEG Working Paper 2007/33, Global Trade Governance Project.

Preeg, Ernest. 1970. *Traders and Diplomats: An Analysis of the Kennedy Round of Negotiations under the GATT*. Brookings Institution.

Putnam, Robert D. 1988. "Diplomacy and Domestic Politics: The Logic of Two-Level Games." *International Organization* 42(3): 427–60.

WTO Secretariat. 2011. *World Trade Report 2011: The WTO and Preferential Trade Agreements: From Co-existence to Coherence*. World Trade Organization.

Journal of Economic Perspectives—Volume 30, Number 1—Winter 2016—Pages 117–138

Will We Ever Stop Using Fossil Fuels?

Thomas Covert, Michael Greenstone, and Christopher R. Knittel

Fossil fuels provide substantial economic benefits, but in recent decades, a series of concerns have arisen about their environmental costs. In the United States, for example, the Clean Air Act in 1970 and 1977 addressed concerns over the emissions of so-called conventional pollutions, notably airborne particulate matter, by imposing fuel economy standards on vehicles and regulations to reduce emissions from stationary sources. During the 1980s, concerns mounted about how the combustion of fossil fuels could lead to acid rain and rising ozone levels. The Clean Air Act Amendments of 1990 created frameworks to reduce sulfur dioxide and nitrogen oxide from power plant emissions, as well as from the combustion of gasoline and diesel fuels in vehicles. However, in many of the world's largest cities in the emerging economies around the world, the conventional forms of air pollution from burning fossil fuels—especially particulates, sulfur oxides, and nitrogen oxides—are still exacting a heavy toll on human health (Chay and

■ *Thomas Covert is Assistant Professor of Microeconomics, Booth School of Business, University of Chicago, Chicago, Illinois. Michael Greenstone is the Milton Friedman Professor in Economics and the College and Director of the Energy Policy Institute at Chicago, both at the University of Chicago, Chicago, Illinois. Christopher R. Knittel is William Barton Rogers Professor of Energy Economics, Sloan School of Management, and Director of the Center for Energy and Environmental Policy Research, all at the Massachusetts Institute of Technology, Cambridge, Massachusetts. Greenstone and Knittel are also Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are thomas.covert@chicagobooth.edu, mgreenst@uchicago.edu, and knittel@mit.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.1.117

doi=10.1257/jep.30.1.117

Greenstone 2003; Chen, Ebenstein, Greenstone, and Li 2013; Knittel, Miller, and Sanders forthcoming).

By the mid-1990s, concerns about the role of fossil fuels in generating emissions of carbon dioxide and other greenhouse gases gained traction. Approximately 65 percent of global greenhouse gas emissions are generated by fossil fuel combustion.[1] Of these emissions, coal is responsible for 45 percent, oil for 35 percent, and natural gas for 20 percent.[2] To reduce carbon dioxide emissions by enough to mitigate the chance of disruptive climate change in a substantial way, there would seem to be only two possible options. One is to find ways to capture carbon from the air and store it. To a moderate extent this can be done by expanding the size of the world's forests, for example, but if carbon capture and storage is to be done at a scale that will more than counterbalance the burning of fossil fuels, there would need to be very dramatic developments in the technologies so that it could happen at a cost-effective industrial level. The other option for reducing emissions of greenhouse gases is to reduce future consumption of fossil fuels in a drastic manner.

A few developed countries have implemented policies to limit fossil fuel consumption through a mixture of taxes, fees, or regulation on carbon emissions (in the case of the European Union, some US states, Japan, Canada, and Australia), subsidies for energy conservation (the United States and elsewhere), and the development of low- or no-carbon energy resources (the United States and elsewhere). In these and other OECD countries, data in the annual *BP Statistical Review of Energy* show that both oil and coal consumption are down about 10 percent, while natural gas consumption (which has lower carbon emissions) is up 10 percent from 2005 levels.[3]

There have been few policy responses to limit fossil fuel consumption in developing countries, even though many of them are experiencing very high and immediate costs from conventional air pollutants. Moreover, this group of countries has greatly expanded its use of fossil fuels in this period, with non-OECD coal, oil, and gas consumption up 46, 33, and 35 percent, respectively, since 2005. As a result, global consumption of fossil fuels rose 7.5 percent for oil, 24 percent for coal, and 20 percent for natural gas from 2005 to 2014.

Two primary market forces could moderate the need for an activist policy response to rising fossil fuel consumption. First, the ongoing consumption of fossil fuels could cause the marginal cost of extracting additional fuel to rise to where the marginal barrel of oil (or ton of coal or cubic meter of natural gas) will be more costly than clean energy technologies (for examples of this argument, see Rutledge

[1] US Environmental Protection Agency at http://www3.epa.gov/climatechange/ghgemissions/global.html.

[2] Carbon Dioxide Information Analysis Center at http://cdiac.ornl.gov/ftp/trends/co2_emis/Preliminary_CO2_emissions_2012.xlsx.

[3] All country-level fuel consumption data in this article come from the *BP Statistical Review of World Energy 2015*, available at http://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html.

2013; Murray and King 2012). We label this "the supply theory"—that is, the world will run out of inexpensive fossil fuels.

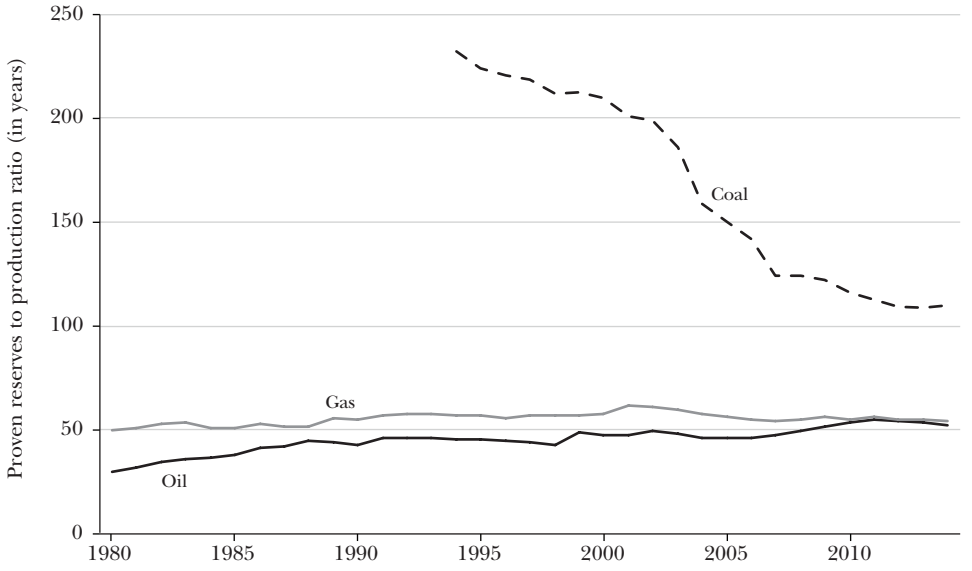
Second, scientific advances could improve the energy efficiency of existing technologies and develop newer, cheaper carbon-free technologies (for example, see the "McKinsey curve" in McKinsey 2009). We label this "the demand theory"—that is, the economy will stop demanding fossil fuels as alternatives become more cost-competitive. This line of thinking is appealing. After all, who wouldn't prefer to consume energy on our current path and gradually switch to cleaner technologies as they become less expensive than fossil fuels? But the desirability of this outcome doesn't assure that it will actually occur—or even that it will be possible.

In this article, we look back at the historical record to assess the power of natural supply-side and demand-side forces (forces unrelated to future policy intervention) to achieve significant reductions in fossil fuel consumption. To gain some initial perspective, consider the ratio of current proven fossil fuel reserves divided by current annual consumption. Proven reserves are defined as fossil fuels that are technically and economically recoverable at current prices, and so this ratio represents the number of years of current consumption that can be economically supplied by known fossil fuel deposits. The total amount of any category of fossil fuels that is in the ground and recoverable *at any cost* is fixed, of course, but reserves can increase or decrease depending on how extraction technologies and prices evolve.

Figure 1 plots reserves-to-consumption for oil, natural gas, and coal at the global level. The available data for oil and natural gas span 1980 to 2014; the data for coal begin in 1994. The striking feature of this graph is how constant the reserve-to-consumption ratio is for both oil and natural gas. It is an empirical regularity that, for both oil and natural gas at any point in the last 30 years, the world has 50 years of reserves in the ground. The corollary, obviously, is that we discover new reserves, each year, roughly equal to that year's consumption. This phenomenon seems to be independent of the enormous variation in fossil fuel price changes over the last 30 years. Coal, on the other hand, shows a dramatic decrease in its reserves-to-production ratio, but there remain many more years of coal reserves, at current consumption, than for either oil or natural gas. Furthermore, the decline in the reserve-to-consumption ratio for coal flattened in the early 2000s and more recently the ratio has even ticked upward a bit. The stability of the reserve-to-consumption ratio provides the first piece of evidence against the idea that the supply of or demand for fossil fuels are likely to "run out" in the medium term.

In the next section of the paper, we analyze the supply behavior for fossil fuels over the past three decades through the lenses of reserve growth and exploration success. The story seems clear: we should not expect the unfettered market to lead to rapid reductions in the supply of fossil fuels. Technical progress in our ability to extract new sources of fossil fuels has marched upward steadily over time. If the advance of technology continues, there is a nearly limitless amount of fossil fuel deposits—at least over the time scale that matters for climate change—that, while they are not yet economical to extract at current prices, could become economical in the future.

Figure 1
Ratio of Proven Fossil Fuel Reserves to Production



Source: BP Statistical Review of World Energy, 2015.

If we cannot rely on market-driven shifts in supply to reduce our consumption of fossil fuels, can we expect the demand for fossil fuels to fall? In the following section, we investigate the prospects for low-carbon alternatives to fossil fuels to become cost-competitive. Hydroelectric, solar, wind, and nuclear are obvious substitutes for fossil fuels in electric power generation. Reducing our demand for petroleum will also require low-carbon sources of transportation, potentially through the large-scale adoption of electric vehicles. We analyze trends in production costs for these cleaner technologies and find it implausible that these trends alone will sufficiently reduce fossil fuel consumption for the world as a whole.

Thus, we are driven to the conclusion that activist and aggressive policy choices are necessary to drive reductions in the consumption of fossil fuels and greenhouse gas emissions. We end the paper with a peek into the future of risks if we were to continue our heavy consumption of known deposits of fossil fuels without capturing and sequestering the emissions. The picture is alarming.

Supply: Peak Oil, Natural Gas, and Coal?

In the aftermath of the oil price shocks of 1973–74, a number of market observers began to warn of ever-dwindling oil resources. A robust debate ensued with geologists and environmentalists often citing Hubbert's (1956) theory of "peak oil" as the basis for the concern. President Jimmy Carter (1977) reflected these

concerns when he told the nation in a televised speech: "World consumption of oil is still going up. If it were possible to keep it rising during the 1970's and 1980's by 5 percent a year, as it has in the past, we could use up all the proven reserves of oil in the entire world by the end of the next decade."

Economists, on the other hand, stressed the ongoing tension between consumption and technological progress. Yes, taken literally, there is a finite amount of any one fossil fuel. Each barrel of oil or cubic foot of natural gas that is taken out of the ground cannot be replaced in a relevant time scale. However, over time we are able to extract more fossil fuels out of the ground as technology improves, and the ultimate resources of planet Earth are both highly uncertain and very large. This observation is perhaps most closely connected with the work of Morris Adelman (1993).

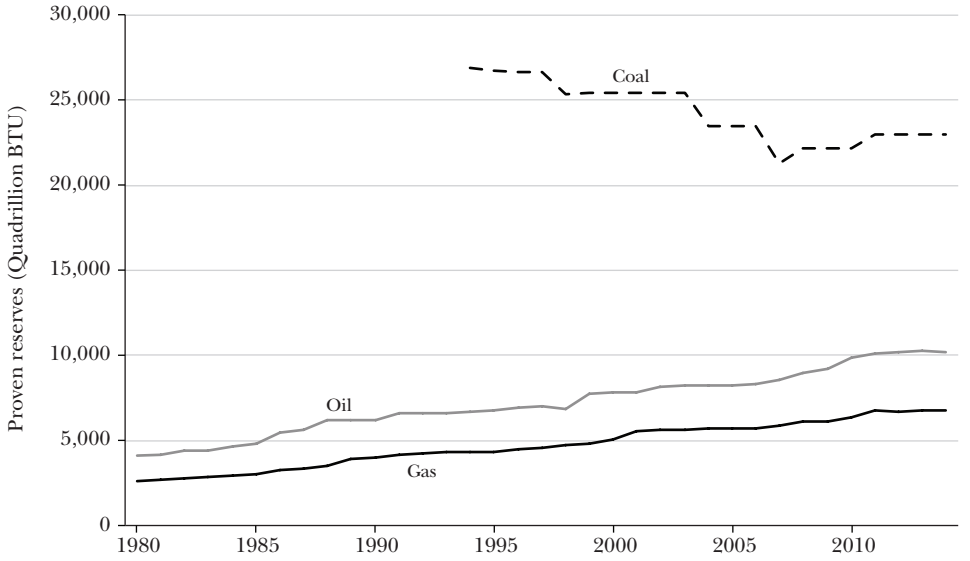
Indeed, two enormous sources of modern oil production—oil from tar sands and oil (as well as gas) from shale deposits—only recently were categorized as "reserves," having previously not been on the radar as economically relevant energy sources. Though these two "unconventional" sources of hydrocarbons currently represent a substantial fraction of total production in the US and Canada and approximately 10 percent of world oil and gas reserves, their economically useful existence was driven entirely by recent technological advances.

Canadian geologists first studied the possibility of production of oil from bituminous sands (also referred to as tar sands or oil sands) in the 19th century (Atkins and MacFadyen 2008). Though the commercial potential for this viscous mixture of heavy oil, sand, and clay had long been recognized, it took scientists several decades to figure out how to economically mine the mixture, separate out the heavy hydrocarbons and "upgrade" them to light, sweet crude oil. It wasn't until 1967 that a small-scale commercial production and upgrading facility began operation (10,000 barrels, hereafter "bbls," per day) and it took until 1999 for Canadian energy authorities to recognize the growing number of tar sands projects as reserves. This decision increased total Canadian oil reserves by 130 billion bbls and total world reserves by about 10 percent. Now, the Canadian Association of Petroleum Producers estimates tar sands production in 2014 was more than 2 million bbls/day.[4]

A similar pattern occurred in the development of oil and gas from shale and other low-permeability rock formations in the United States. These resources had long been known to contain tremendous quantities of hydrocarbons (Shellberger, Nordhaus, Trembath, and Jenkins 2012). However, their low permeability inhibited the rate at which oil and gas molecules could flow out of conventionally designed vertical wells drilled into them. In the 1980s, engineers working in the Barnett natural gas shale formation in Texas began experimenting with hydraulic fracturing and horizontal drilling as tools to solve the permeability problem. By the early 2000s, thousands of wells had been successfully drilled and "fracked" into the Barnett, and the technique

[4] This figure includes synthetic "upgraded" crude oil as well as raw bitumen production, according to data from the Canadian Association of Petroleum Producers at http://statshbnew.capp.ca/SHB/Sheet.asp?SectionID=3&SheetID=233.

Figure 2
Proven Reserves of Oil, Natural Gas, and Coal over Time



Source: BP Statistical Review of World Energy, 2015.
Note: The figure plots annual reserve estimates in absolute terms for oil, natural gas, and coal.

later spread to the Marcellus natural gas shale formation in Pennsylvania and the Bakken oil shale formation in North Dakota. As a result, US oil and gas reserves expanded 59 and 94 percent, respectively, between 2000 and 2014. This technology-driven growth in reserves also caused increases in production. In 2014, US natural gas production was greater than ever before and oil production was at 97 percent of the peak reached in 1970.

These two technological advances are at least partially responsible for a more general long-term pattern of worldwide reserve growth. Figure 2 plots annual reserve estimates in absolute terms for oil, natural gas, and coal. The steady rise of proven oil and natural gas reserves is striking. The average growth rate in reserves is 2.7 percent for both oil and natural gas. In only one year did total proven oil reserves fall, and this year was immediately followed by a growth in reserves of 12.2 percent the following year. Natural gas reserves fell in only two years, but by less than half of 1 percent in both cases. Coal reserves, on the other hand, fell consistently through the late 1990s to 2008 but have since shown fairly constant positive growth.

A potential concern about the consistency of this pattern is that individual countries do not consistently scale back reserve estimates when low prices cause existing discoveries to be unprofitable. While there is no uniform standard that all countries use in calculating reserves, there are at least two reasons to believe that these data are informative about the scale of fossil fuel resources readily available in the future. First, securities regulators in developed countries heavily regulate

reserve estimates published by publicly traded oil and gas companies, and these numbers are regularly revised downward following oil price crashes. For example, in 1986 and 1998, when oil prices fell 46 and 30 percent, respectively, US oil reserves fell by 3 and 6 percent, respectively. Second, even in the absence of truth-telling regulation in reserve estimates (for example, in the case of oil owned by national oil companies), the country-level data in the *BP Statistical Review* still shows a meaningful number of downward revisions in a typical year. In the history of oil reserve changes in the *BP Statistical Review*, a full quarter of countries report decreases in reserve estimates in the average year. However, the converse of this statement is what is important for understanding how hydrocarbon reserves may change in the future. The BP data show that most countries seem to find significantly more oil (and gas) than they consume in most years.

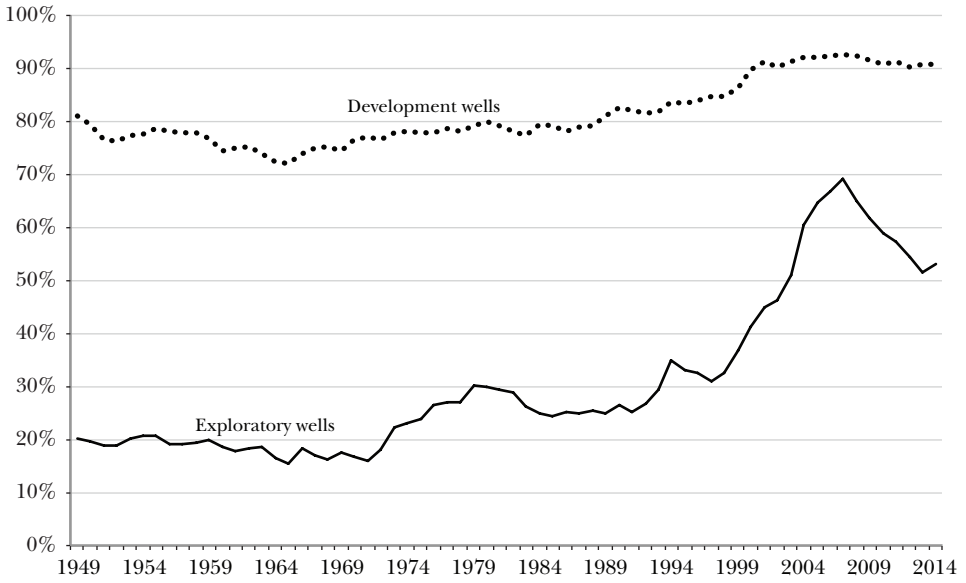
If the past 35 years is any guide, not only should we not expect to run out of fossil fuels any time soon, we should not expect to have less fossil fuels in the future than we do now. In short, the world is likely to be awash in fossil fuels for decades and perhaps even centuries to come.

An alternative measure of technological progress in the exploration and exploitation of fossil fuel resources is the success rate associated with exploring for new oil and natural gas formations. Without technological progress in this area, present-value considerations will cause firms to drill lower-risk prospects before higher-risk prospects. As a result, the probability of successful exploration falls over time as the "easy" wells are exhausted and oil and gas prices rise to equilibrate supply and demand. In contrast, if technology advances and lowers exploration costs and/or risks, it is possible for the probability of successful exploration to stay constant, or even rise over time, independent of the path of prices.

We are unaware of any systematic data on the history of exploration and development costs for oil or gas, given that it is difficult to observe private company costs for seismic studies, drilling, and other inputs to the exploration process. However, the US Energy Information Administration and the data and consulting firm IHS do publish data on the number of successful and failed exploration and development wells in the United States, so it is possible to measure changes in risk over time. Figure 3 plots the fractions of successful exploration and development wells in each year from 1949 to 2014. The probability of a successful exploratory well did drift downward from about 20 percent in 1949 to 16 percent in the late 1960s. But in 1968, the highly successful Alaska North Slope field was discovered, leading to a near doubling in the probability of exploratory drilling success by 1979.

Similar technological events preceded other periods of growing drilling success. These include the development of ultra-deep water fields in the Gulf of Mexico during the 1980s, hydraulic fracturing technology for natural gas formations in the 1990s, and the same technology for oil formations in the 2000s. By 2007, 69 percent of exploratory wells yielded successful oil or gas production. Though the probability of successful exploration has drifted down to about 50 percent in recent years, it is still markedly higher than in much of the history of US fossil fuel exploration. Figure 3 also shows that the fraction of successful development wells has also

Figure 3
Fraction of Development and Exploratory Wells that are Successful in the United States



Source: US Energy Information Administration and IHS.
Notes: Development wells are drilled into formations that have already been explored and thus are known to contain oil or gas. Exploratory wells are drilled into formations that have not yet been explored and which might not contain oil or gas.

grown over time. Though these wells are drilled into formations already known to contain oil or gas, there is still risk that a development well faces technical difficulties and produces no output. Growth in this number is also important, since there have been 10 to 20 times as many development wells compared to exploratory wells in recent years. In the United States, at least, it appears that technical progress has consistently helped increase the supply of fossil fuels, in spite of price volatility and the exhaustion of existing fossil fuel formations.

The supply curve for fossil fuels has constantly shifted out over large stretches of time, both because of new discoveries (like the large-scale development of new oil sources or oil and gas from Alaska and from the North Sea in the 1970s) as well as from new techniques like deep-water drilling, hydraulic fracturing, and extracting oil from tar sands. What might be next on the horizon?

Besides measuring fossil fuel reserves, geologists also measure fossil fuel "resources"—that is, fossil fuel deposits that are known to exist but are not currently economical to extract. McGlade and Ekins (2015) summarize reported resources by the Federal Institute for Geosciences and Natural Resources, the International Energy Agency, and the Global Energy Assessment. The range of oil resources is from 4.2 to 6.0 trillion barrels, which is 2.8 to 4 times larger than the existing

reserves of roughly 1.5 trillion barrels. The range of natural gas resources is also immense, ranging from 28,000 trillion cubic feet to over 410,000 trillion cubic feet. For comparison, current global reserves of natural gas are roughly 7,000 trillion cubic feet (according to the US Energy Information Administration's International Energy Statistics).[5] Finally, the estimate range for coal resources is from 14,000 to 23,500 gigatons, compared to existing reserves of around 1,000 gigatons.

These existing "resources" for currently exploited fossil fuel technologies could potentially, if technological progress continues to advance, allow the supply curves for fossil fuels to continue to shift outward for quite some time. In addition, two notable additional resources have not yet been commercially developed, but are known to exist in large quantities: oil shale and methane hydrates.

"Oil shale" is defined by the US Geological Survey as "fine-grained sedimentary rock containing organic matter that yields substantial amounts of oil and combustible gas upon destructive distillation" (Dyni 2006). Despite the similarity in nomenclature, "oil shale" is fundamentally different from the earlier-mentioned "shale oil" which is currently being extracted using hydraulic fracturing techniques in North Dakota, Texas, and Oklahoma. The production technology for oil shale is similar to that of oil sands. In both cases, a heavy hydrocarbon (bitumen in the case of oil sands and kerogen in the case of oil shale) exists in sand, clay, or sedimentary rock. Heat is used to separate the hydrocarbon from the surrounding material. The resulting bitumen or kerogen then goes through additional refining steps to become a final consumer product. Oil shale resources are enormous. A US Geological Survey report from 2006 estimates that 2.8 trillion barrels of oil shale exist (that is, almost 50 percent of the high-end estimate of oil resources), and some private estimates are much larger. If oil shale became economical in the near future, it could cause oil reserves to nearly triple. Because of the use of heat in the extraction phase and the extra steps required for refining, carbon emissions for producing oil from oil shale are greater than for conventional sources of oil. Some estimates suggest that the emissions are 21–47 percent greater per unit of energy produced compared to conventional sources (Brandt 2008).

Methane hydrates are a solid mixture of natural gas and water that forms in low-temperature and high-pressure environments, usually beneath seafloors. Geologists have recognized methane hydrates as a potential source of hydrocarbons since the 1960s, and testing on whether the resource might be commercially viable dates back to the 1970s. The technology to extract methane hydrates at a commercial price does not currently exist, although a number of countries are actively pursuing technology in this area (Boswell et al. 2014). If this technology eventually achieves commercial success, the potential scale of methane hydrate resources ranges from 10,000 trillion cubic feet to more than 100,000 trillion cubic feet, according to US Geological Survey estimates (US Energy Information

[5] The page of the International Energy Statistics on which this data is found is: http://www.eia.gov/cfapps/ipdbproject/iedindex3.cfm?tid=3&pid=3&aid=6&cid=ww,&syid=2011&eyid=2015&unit=TCF.

Administration 2012). For comparison, current global reserves of natural gas are roughly 7,000 trillion cubic feet.

The status of oil shale and methane hydrates today is similar to that of oil sands and shale gas in the 1980s. Geologists knew of their existence, but oil and gas companies did not yet know how to recover them in a cost-effective manner. The remarkably successful history of innovation in oil and gas exploration makes it seem more than possible that oil shale and methane hydrates will become commercially developed.

Even without the emergence of new technology for oil shale and methane hydrates, the use of existing technology to develop shale gas and shale oil resources outside of the United States is just now starting. The US Energy Information Administration (2013) estimates that 93 percent of shale oil and 90 percent of shale gas resources exist outside of the United States. Further, these resources represent 10 and 32 percent, respectively, of total world resources for oil and gas. As shale technology spreads across the world, these resources are likely to become economically productive reserves.

The policy implications of this ongoing expansion of fossil fuel resources are potentially profound. Even if countries were to enact policies that raised the cost of fossil fuels, like a carbon tax or a cap-and-trade system for carbon emissions, history suggests that technology will work in the opposite direction by reducing the costs of extracting fossil fuels and shifting their supply curves out.

Demand: Will Low-Carbon Energy Sources Knock Fossil Fuels Out of the Money?

Absent large upward shifts of the supply curve for fossil fuels, deep cuts in fossil fuel consumption will have to come from inward shifts in their respective demand curves.[6] In this section, we analyze the recent changes in the relative prices of carbon-free and fossil-fuel–based energy technologies and characterize what future relative prices are necessary to reduce demand for fossil fuels in a substantial manner. We focus on the electricity and transportation sectors, which are major users of fossil fuels. In the United States, for example, the electricity sector consumes over 90 percent of total coal consumption and 30 percent of natural gas, while the transportation sector consumes over 60 percent of total US oil consumption.[7]

[6] We admittedly are confounding shifts in the supply curve with changes in the slope. That is, a shift would represent the addition of zero-marginal-cost supplies. In practice, the marginal cost of many of the new resources is nonzero, implying that the supply curve shifts, or becomes flatter, at some strictly positive price.

[7] US statistics on the end-uses of fossil fuels are readily available at the website of the US Energy Information Administration. For example, coal statistics are available at from the *Quarterly Coal Report* at http://www.eia.gov/coal/production/quarterly; natural gas statistics from the "Natural Gas Consumption by End-Use" page at http://www.eia.gov/dnav/ng/ng_cons_sum_dcu_nus_a.htm; and oil statistics at the "Petroleum and Other Liquids" webpage at http://www.eia.gov/dnav/pet/pet_cons_psup_dc_nus_mbbl_a.htm.

Replacing Natural Gas and Coal in Electricity Generation

Solar photovoltaics, wind turbines, and nuclear fission power plants are the current leading candidates to replace coal and natural gas in electricity generation. However, nuclear fission has already been commercially exploited for almost 75 years. In spite of this maturity, the rate of new nuclear power plant construction has significantly slowed down since the 1980s, and its share of power generation in most countries has actually fallen over the last decade due to decreasing cost competitiveness (Deutch et al. 2009; in this journal, Davis 2012; see also US EIA International Energy Statistics). Thus, we focus here on the recent history of cost improvements in solar photovoltaics and wind turbines.

We compare costs using *levelized cost of energy* estimates across different technologies. The levelized cost of energy is the present discounted value of costs associated with an energy technology divided by the present discounted value of production—that is, it is a measure of the long-run average cost of the energy source.[8] This measure offers a way of adjusting for the fact that renewable energy sources and fossil fuel plants have a different profile of costs over time. For both renewable plants and fossil fuel plants, the single largest cost occurs in the first year a plant is built, representing the up-front capital cost of the plant. However, for renewable technologies, the operating costs in the remaining years are small, as there are no fuel inputs necessary, only maintenance. In contrast, fossil fuel technologies like natural gas and coal require ongoing fuel costs. Thus, a comparison between the levelized cost of energy for a renewable energy technology and a fossil fuel technology hinges on the difference in up-front capital costs between the two technologies relative to expected fuel costs for the fossil fuel technology.

We report forecasts of the levelized cost of electricity generation published by the US Energy Information Administration during the last two decades. For each year in our data, these forecasts report a projected levelized cost of energy for coal, natural gas, nuclear, and wind for electricity generation plants to be built 5–10 years in the future. The forecast for solar is on a similar timeframe, but for a specific year (usually 4–7 years in the future).[9] We prefer this set of estimates of levelized costs of energy to others that are available for a number of reasons: they are produced annually; the methodology is clearly explained and documented; a wide range of electricity-generating technologies is compared; and the time series is relatively long.

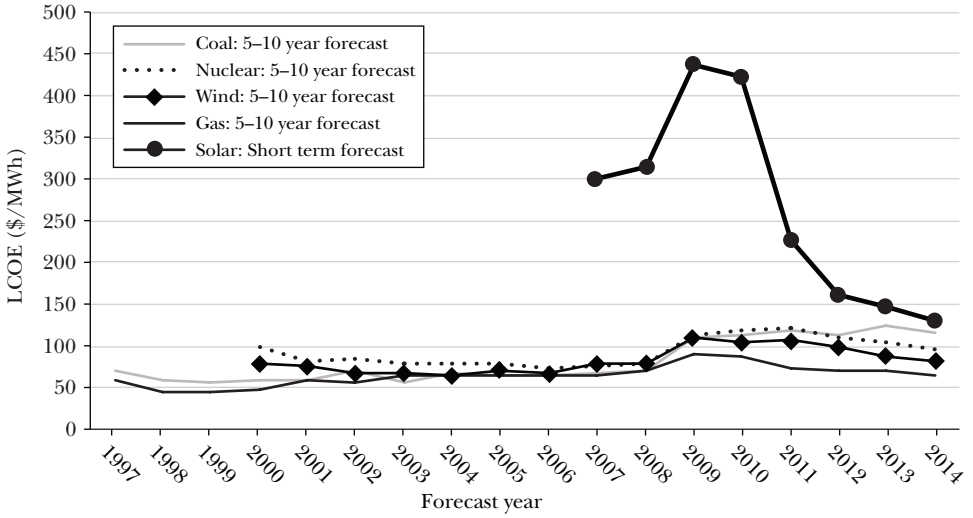
[8] Specifically, the levelized cost of energy cost of energy is calculated as:

$$LCOE = \frac{\sum_{t=0}^{T} \delta^t C_t}{\sum_{t=0}^{T} \delta^t q_t},$$

where δ is the discount factor, T is the lifetime of a generating technology, and C and q are the technology's per-period costs and production, respectively.

[9] For 2007, the projection is for 2014; for 2008 through 2011, it is for 2016; for 2012 it is for 2017; 2013 for 2018; and, 2014 for 2019. Therefore, the average across all of these years is a six-year forecast, while the last four years report a five-year forecast.

Figure 4
Levelized Cost of Energy (LCOE) Forecasts from the US Energy Information Administration



Source: EIA Annual Energy Outlook reports from 1997 to 2014.

One could certainly argue that a longer-term cost outlook would be more appropriate. We focus on estimates 5–10 years out for at least two reasons. First, many analysts believe that it is important to reduce CO_2 emissions in the next decade to mitigate the odds of disruptive climate change. Second, while longer-term costs estimates might exist (although we are unaware of a reasonably long series), they are believed to be very imprecise.

Figure 4 plots the forecasts from the US Energy Information Administration for the last 18 years. Perhaps the most striking finding shown in the figure is the dramatic fall in the cost of solar energy during the past five years. The 2009 forecast for the near-term levelized cost of solar power was nearly $450/MWh of electricity generated, while the 2014 number is under $150/MWh. The speed of these reductions appears to have subsided, but the downward trend continues.

The decline in the costs of solar energy is indeed rapid, but it seems plausible. For example, the current cost of solar energy can be inferred from auctions of photovoltaic installations—in which the prices imply long-run average cost per megawatt-hour of electricity generated. For example, in November 2014, Dubai's state utility held an auction for 100 MWs of photovoltaic power over a 25-year period. The lowest bid was $59.8/MWh. More generally, Bollinger and Seel (2015) document that utilities in the southwestern parts of the United States are now routinely acquiring power from new solar photovoltaic projects at prices in the range of $40–$50/MWh. Though these new projects receive a federal investment tax credit

equal to 30 percent of construction cost, their implied "real" costs are still near or below the levelized cost of energy estimates for natural gas generation in Figure 4.

However, these examples of highly competitive prices for solar energy are from specific locations that are exceptionally well-suited for generating solar energy. In contrast, Figure 4 reports the *average* forecast across locations, including locations where solar exposure may be limited or wind speeds may be slow. Thus, the *average* levelized cost of solar—even given that more favorable sites are being chosen for solar—is still twice that of natural gas in the United States. Wind power seems close to cost competitive with fossil fuel generation in many locations. As with solar, it is complicated to calculate the exact underlying costs of wind power, due to the availability of investment and production tax credits and support from state renewable portfolio standards.

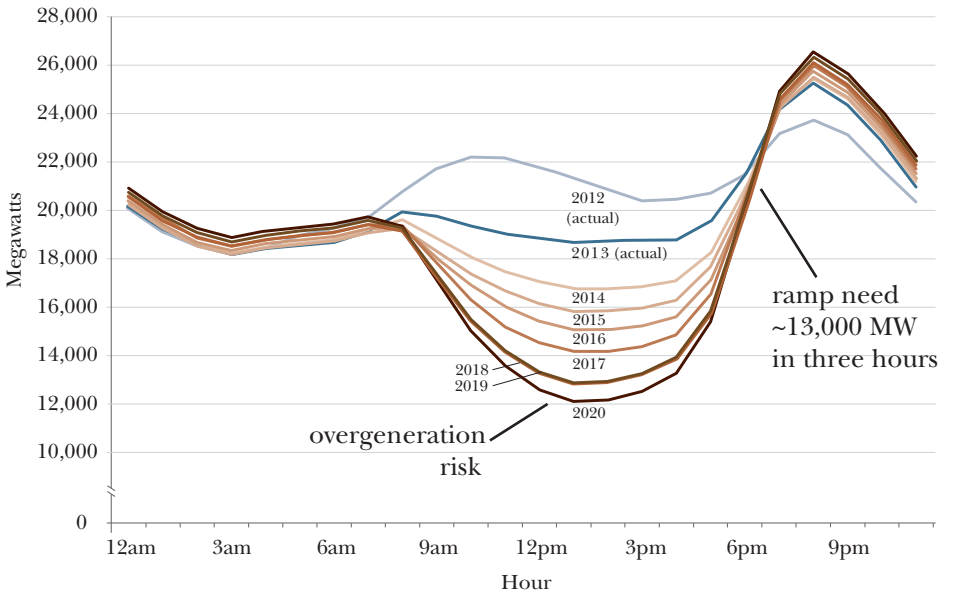
The levelized cost for an electricity technology is only one dimension to be taken into account in making comparisons. Three additional challenges exist. First, both the intensity of sunlight and the speed of the wind vary tremendously across space,[10] meaning that the same solar panel or wind turbine installed in one location will generate vastly different amounts of electricity than if it were installed in another location. This implies that the long-run marginal cost of solar and wind will be upward sloping and large-scale deployment in some areas is likely to be infeasible.

Second, solar and wind energy are inevitably intermittent, which requires either increases in backup generation (often supplied by natural gas generators) or increases in energy storage that aren't typically reflected in the numerator of the basic levelized cost of energy calculation. These costs will depend upon a variety of factors, such as the level of penetration, the degree of variation in generation from the renewable resources, and the correlation in generation across renewable resources. While more research on magnitude of these costs is needed, some estimates exist; these estimates are likely to be very site-specific since they depend on the variability in solar and wind availability. The bulk of this research simulates electricity systems under varying penetration levels of renewables assuming a specific location. Mills and Wiser (2010) simulate the costs of operating an electricity system under different penetration of renewables using detailed simulated output from Midwestern solar and wind installations. They find that for a 10 percent penetration of solar photovoltaic power, intermittency can add as much as $39 per MWh, when the solar is installed in one location, to as little as $3 per MWh when it is installed across 25 different sites. The intermittency costs of a 10 percent penetration of wind across 25 sites are simulated to be below $2 per MWh. In contrast, Wolak (2015) suggests that the costs of intermittency in California are likely to be high because the benefits from diversifying site locations are small.

Finally, because the generation from solar (and wind) resources in a given area tend to be positively correlated, large-scale penetration of either resource will

[10] It is also possible for fossil fuel technologies to have different localized cost of energy in different places as a result of fuel transportation constraints. For example, natural gas prices and coal prices vary significantly across the United States (as shown by US Energy Information Administration data).

Figure 5
California ISO's "Duck Curve"
(Net load–March 31)



Source: Figure 2 from "Fast Facts" published by California ISO, available at: https://www.caiso.com/Documents/FlexibleResourcesHelpRenewables_FastFacts.pdf.
Notes: This graph shows net electricity demand (load), across the hours of the day on March 31, as California approaches penetration of 30 percent renewables by 2020. Because the renewables will predominately be solar, and solar generation peaks during the day, net demand will continue to fall during the day.

inevitably reduce the value of incremental capacity additions. The impact of this on net demand for electricity (after netting out the supply of solar resources) within California has generated what is being referred to as the California Independent System Operator's (CAISO) "Duck Curve," represented in Figure 5. The CAISO has forecast demand for electricity, net of renewable generation, in each year through 2020, when the required amount of generation from renewables hits 30 percent. Figure 5 shows these forecasts for March 31. The figure illustrates that as progressively more and more renewables hit the market, net demand will be lowest during daytime hours and prices during those hours will obviously fall, making additional investments in renewables less valuable.

Intermittency and the large reductions in net demand during peak generation periods imply that, absent economical storage technologies, solar and wind power are ill-suited for baseload generation which is currently covered by coal, natural gas, nuclear, and hydroelectric power.

The levelized costs for fossil fuel technologies presented above also ignore externalities. How much would pricing the externalities associated with carbon

dioxide emissions change these conclusions? Based on Greenstone, Kopits, and Wolverton (2013), the US government applies a social cost of carbon of $43 per metric ton of carbon dioxide in 2015 dollars. Using this value, the carbon dioxide externality for the typical natural gas plant is $20 per MWh, while the externality for the typical coal plant is $40 per MWh. But the US Energy Information Administration forecasts that the gap between average levelized costs for solar and natural gas will still be about $50 per MWh in the near-term.

Replacing Oil Usage in Motor Vehicles

While there are many substitutes for fossil fuels in electricity generation, the primary path for moving away from them in the transportation sector is the use of battery powered electric vehicles, which in turn requires several technology breakthroughs to occur. First, even if oil prices were at $100 per barrel, the price of batteries that store the energy necessary to power these vehicles needs to decrease by a factor of three. Second, the time needed for these batteries to charge must be shortened. Third, the electricity that is fueling these cars will need to have sufficiently lower carbon content than petroleum. Otherwise, we could transition from oil-based transportation with moderately high carbon emissions to coal-fired-electricity-based transportation with even higher carbon emissions. As noted in Graff Zivin, Kotchen, and Mansur (2014), effective carbon emissions from electric vehicles that are powered by the existing US power plant fleet are generally higher than emissions from high-efficiency gasoline-powered vehicles. Only 12 percent of fossil-fueled power plants have low enough carbon emissions that electric vehicles powered by them would have lower emissions than a Toyota Prius.

The large-scale adoption of electric vehicles, instead of petroleum-based internal combustion engine vehicles, seems likely to require all three of these events. To date, none of them have happened. The previous subsection discussed the prospects for greening the electrical grid. Refueling times, as well as the absence of an abundant refueling infrastructure, remain challenges. First we will explore the necessary innovation in batteries.

What is the magnitude of the necessary improvement in storage technologies? Here, we describe some back-of-the-envelope calculations that provide a sense of the present discounted value of operating an internal combustion engine, compared with the present discounted cost of operating an electric vehicle. The bottom line from these calculations is that truly dramatic improvements in battery technology are necessary to bring these technologies into cost parity.

We want to compare the operating costs of an electric vehicle to an internal combustion engine. We assume a 3,000-pound vehicle. (For comparison, the 2015 Honda Accord has four doors and a four-cylinder engine, weighs about 3,200 pounds, and gets a combined-fuel economy of 33 miles per gallon.[11]) We

[11] The official fuel economy rating is 31 miles per gallon, while the user-reported fuel economy is 33.8 miles per gallon. See the US Department of Energy website at https://www.fueleconomy.gov/feg/PowerSearch.do?action=noform&path=7&year=2015&make=Honda&model=Accord&srchtyp=ymm.

assume both the electric car and the internal combustion engine car are driven 15,000 miles per year, and we use a discount rate of 5 percent.

For the internal combustion engine, we assume that it presently gets 30 miles per gallon. For the electric vehicle, we consider a battery size for a range of 250 miles, which of course is still shorter than driving range of most current internal combustion vehicles. We use a price for purchasing electricity of 12.2 cents per kWh, which is consistent with the average US retail price in 2014. The electric vehicle is assumed to consume 0.3 kilowatt-hours electricity per mile. Finally, because internal combustion engines tend to be more costly than electric motors, we "credit" electric vehicles by $1,000 (Peterson and Michalek 2013).

Current battery costs for an electric vehicle are roughly $325 per kWh. This estimate is consistent with the cost of Tesla's Powerwall home battery, which retails for a price of $350 per kWh for the 10 kWh model (and does not include the price of an inverter for use in the home). This cost estimate may be lower than an average battery cost. For example, Tesla charges $3,000 for an extra 5 kWhs of battery capacity in its Model S, which is $600 per extra kWh; however, this incremental price may also include some level of price discrimination on the part of Tesla. In a 2014 "EV Everywhere Grand Challenge" study, the US Department of Energy finds that the current battery cost is $325 per kWh.[12] At a battery cost of $325 per kWh, the price of oil would need to exceed $350 per barrel before the electric vehicle was cheaper to operate.[13] During 2015, the average price of oil was approximately $49 per barrel. At present, the costs of batteries make large-scale penetration of electric vehicles unlikely.

How is this comparison between the operating costs of an internal combustion engine and an electric vehicle likely to evolve into the future? In these kinds of comparisons, a common mistake is to compare *future* costs of an electric vehicle with *current* costs of an internal combustion engine. But this is the wrong test. Future electric vehicles will be competing against future combustion engine vehicles, not current ones. There will be technological progress in all areas, not only for low-carbon technologies.

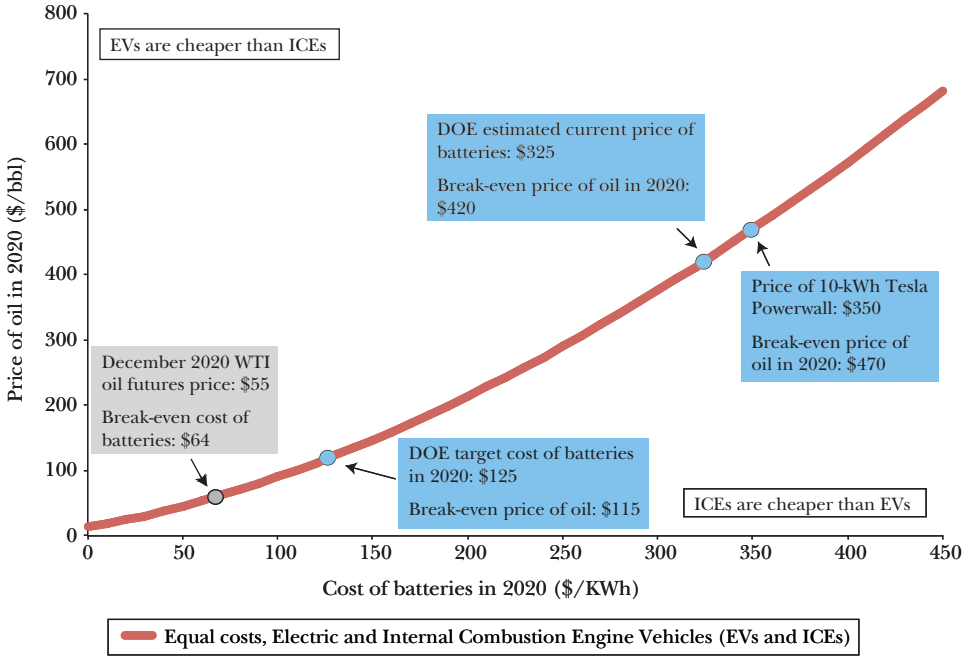
The historical record suggests that we should expect persistent innovation in the efficiency of combustion engine vehicles, just as it suggests that there will be continued innovation in the extraction of fossil fuels. For the internal combustion engine, we assume that the fuel economy grows at 2 percent per year, which is consistent with Knittel (2011).

As a basis for estimating the future price trajectory of batteries, Nykvist and Nilsson (2015) survey 85 peer-reviewed estimates of current and future battery costs published since 2008. They also report battery cost estimates for the Nissan Leaf,

[12] A presentation connected to the study gives a range of $325–500 per kWh. http://energy.gov/sites/prod/files/2014/03/f8/5_howell_b.pdf.

[13] To connect the price of gasoline to the price of oil, we regress the log of historic gasoline prices on the log of oil prices to capture the nonlinear relationship between gasoline and oil prices. We use data from the US Energy Information Administration from April 1993 to July 2015. The estimated intercept is −1.57; the slope is 0.6046. The R^2 is 0.98. The general conclusions are unchanged if we instead assume a linear relationship between gasoline and oil prices.

Figure 6
Break-even Oil and Battery Costs



Notes: This graph plots the relationship between oil prices and battery costs such that the present discounted value of owning and operating an internal combustion engine vehicle equals the present discounted value of owning and operating an electric vehicle assuming the year is 2020. We assume the vehicle is driven 15,000 miles per year. We use the average retail price of electricity of 12.2 cents per kWh to charge the electric vehicle. We estimate the relationship between oil prices and gasoline prices using monthly data from 1993 to 2015 and assume a log–log relationship. We size the battery such that the electric vehicle has a range of 250 miles, assuming electricity consumption of 0.3 kWh per mile. The cost of the electric vehicle is reduced by $1,000 to reflect the lower costs associated with electric motors relative to the internal combustion engine vehicle. The fuel economy of internal combustion engines is assumed to grow at an annual rate of 2 percent, consistent with Knittel (2011).

Tesla S, and other electric vehicles. They find large reductions in battery costs over the past 10 years. However, their data also suggest battery costs are predicted to level off between $150 to $300 per kWh over the next 15 years. The same US Department of Energy (2014) "EV Everywhere" presentation discussed above defines a "target" battery cost of $125 per kWh by 2022.

We can extend the calculation above for a variety of oil-price–battery-cost pairs by calculating the indifference price of oil for a given battery cost. Figure 6 plots this relationship across battery costs of 0 to $400 per kWh. For this graph, we compare the technologies in 2020 allowing for the efficiency of internal combustion engine vehicles to increase by 2 percent per year. Even at the US Department of Energy target price for 2020, oil prices would have to rise to $115 per barrel for electric

vehicles to be cost-competitive with internal combustion engines under the assumptions discussed above. If battery costs remain at $325 per kWh, oil prices would have to exceed $420 per barrel. For comparison, the current December 2020 oil futures price using the West Texas Intermediate benchmark (observed on December 18, 2015) was $55/barrel, requiring a battery cost that would fall to $64 per kWh.

These basic calculations make it clear that at least for the next decade or two, electric vehicles face an uphill battle. Not only are large continuing decreases in the price of batteries necessary, but oil prices would have to increase by more than financial markets currently predict.

Other barriers to widespread electric vehicle adoption are not reflected in these calculations. First, the estimates are built on a battery with a range of 250 miles, which for some drivers would not be enough. Second, we assume that there is no disutility associated with longer recharging times of electric vehicles relative to fueling times for internal combustion engines. Finally, oil prices are endogenous, so substantial penetration of electric vehicles would reduce demand for oil. Provided that the supply curve for oil is upward sloping (as it is in almost all markets), this drop in demand would translate into lower oil prices, making gasoline vehicles more attractive.

How does this comparison change if we include the social cost of carbon dioxide emissions? Assume that 20 pounds of carbon dioxide is emitted into the atmosphere for every gallon of gasoline burned. (Most of the weight of the carbon dioxide arises when the car emissions of carbon atoms in the gasoline combine with oxygen in the air, which is why a gallon of gasoline weighing about 6 pounds can produce 20 pounds of carbon dioxide.) As we noted above, a battery price of $125 kWh implies $115/barrel of oil as a break-even price for an internal combustion engine. If we account for the social cost of carbon and make the extreme assumption that electricity for the electric vehicle is carbon free, the break-even price for oil falls to $90 per barrel.

Our emphasis in this section has been on the scope for battery technologies to unseat oil in transportation as well as solar and wind resources as noncarbon methods of generating electricity. There are other noncarbon alternatives for producing energy. In some parts of the world, including Africa and South America, some proportion of the rising demand for electricity might be met by hydroelectric power. Certain regions may have possibilities for electricity generated by geothermal energy or ocean thermal gradients. We mentioned earlier that we were setting aside any discussion of nuclear power in this paper, given the high and rapidly rising costs of nuclear power generation in recent decades. We are generally supportive of research and development into all of these other noncarbon methods of generating electricity. However, the International Energy Administration Agency (2015) projects that fossil fuels will account for 79 percent of total energy supply in 2040 under the current, business-as-usual policies, which already takes into account some rise in these alternative noncarbon energy production technologies. In the medium-run of the next few decades, none of these alternatives seem to have the potential based on their production costs (that is, without government policies to raise the costs of carbon emissions) to reduce the use of fossil fuels dramatically below these projections.

Discussion

Our conclusion is that in the absence of substantial greenhouse gas policies, the US and the global economy are unlikely to stop relying on fossil fuels as the primary source of energy. The physical supply of fossil fuels is highly unlikely to run out, especially if future technological change makes major new sources like oil shale and methane hydrates commercially viable. Alternative sources of clean energy like solar and wind power, which can be used both to generate electricity and to fuel electric vehicles, have seen substantial progress in reducing costs, but at least in the short- and middle-term, they are unlikely to play a major role in base-load electrical capacity or in replacing petroleum-fueled internal combustion engines. Thus, the current, business-as-usual combination of markets and policies doesn't seem likely to diminish greenhouse gases on their own.

What are the consequences of a continued reliance on fossil fuels? We conducted some back-of-the-envelope calculations of the potential warming associated with using all available fossil fuels. This requires estimates of total reserves and resources of each fossil fuel, carbon conversion factors, estimates of historical emissions, and a model to convert carbon dioxide emissions into temperature changes. It is important to note that this exercise is based on high levels of greenhouse gas emissions for many decades beyond 2100. For example, our calculations are based on total carbon emissions ranging from 12,744 to 17,407 gigatons of CO_2. For comparison, the business-as-usual scenario from IPCC (2013) has cumulative emissions of 6,180 gigatons of CO_2 between 2012 and 2100. The calculations are described in detail in a web appendix.[14] Our headline finding is that the combustion of currently known fossil fuels would increase global average temperatures by 10°F to 15°F, depending on the choice of carbon conversion factors and model. Such scenarios imply difficult-to-imagine change in the planet and dramatic threats to human well-being in many parts of the world. Further, these estimates do not account for advances in fossil fuel extraction techniques that could make other deposits economically accessible; for example, the use of oil shale and methane hydrate deposits would add another 1.5°F to 6.2°F of warming. Research on the economic consequences of such changes in temperatures is an important area that is rapidly advancing (for example, Deschênes and Greenstone 2011).

What are the prospects for avoiding this dystopian future? At a high level, there are two market failures—greenhouse gas emissions are not priced adequately, and basic or appropriable research and development is too often underfunded—and the corresponding solutions of pricing emissions and subsidizing basic research and development are easy to identify. However, the politics of implementing such policies are complex, particularly because energy consumption is projected to grow the most in low- and middle-income countries during the coming

[14] This appendix is available with this paper at http://e-jep.org. Alternatively, see the following link for details on these calculations: https://epic.uchicago.edu/sites/default/files/Calculating%20the%20Temperature%20Potential%20of%20Fossil%20Fuels.pdf.

decades, and thus the majority of emissions cuts would need to take place in those countries. Without direct compensation, global emissions cuts will require the poorer countries to use and pay for more expensive energy sources. The last several decades have seen limited global progress in tackling these policy problems.

The Conference of Parties (COP21) climate conference in Paris in December 2015 has set out the broad outlines of what could constitute a dramatic change in global climate policy. Whether this high-level voluntary agreement leads to the climate policies necessary to correct the market failures related to greenhouse gases around the globe will be determined in the coming years and decades. Ultimately, their enactment would greatly reduce the probability that the world will have to contend with disruptive climate change. The alternative is to hope that the fickle finger of fate will point the way to low-carbon energy sources that rapidly become cheaper than the abundant fossil fuels on their own. But hope is too infrequently a successful strategy.

References

Adelman, Morris A. 1993. *The Economics of Petroleum Supply: Papers by M. A. Adelman, 1962–1993.* MIT Press.

Atkins, Frank J., and Alan J. MacFadyen. 2008. "A Resource Whose Time Has Come? The Alberta Oil Sands as an Economic Resource." *Energy Journal* 29(special issue): 77–98.

Bolinger, Brian, and Joachim Seel. 2015. "Utility-Scale Solar 2014: An Empirical Analysis of Project Cost, Performance, and Pricing Trends in the United States." Lawrence Berkeley National Laboratory report LBNL-1000917.

Boswell, Ray, Koji Yamamoto, Sung-Rock Lee, Timothy Collett, Pushpendra Kumar, and Scott Dallimore. 2014. "Methane Hydrates." Chap. 8 in *Future Energy*, 2nd edition. Elsevier.

BP. 2015. *BP Statistical Review of World Energy 2015.* Available at: http://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html.

Brandt, Adam R. 2008. "Converting Oil Shale to Liquid Fuels: Energy Inputs and Greenhouse Gas Emissions of the Shell in Situ Conversion Process." *Environmental Science and Technology* 42(19): 7489–95.

Carbon Dioxide Information Analysis Center. No date. http://cdiac.ornl.gov/ftp/trends/co2_emis/Preliminary_CO2_emissions_2012.xlsx.

Carter, Jimmy. 1977. "Address to the Nation on Energy." April 18. The American Presidency Project. http://www.presidency.ucsb.edu/ws/index.php?pid=7369.

Chay, Kenneth Y., and Michael Greenstone. 2003. "The Impact of Air Pollution on Infant Mortality: Evidence from Geographic Variation in Pollution Shocks Induced by a Recession." *Quarterly Journal of Economics* 118(3): 1121–67.

Chen, Yuyu, Avraham Ebenstein, Michael Greenstone, and Hongbin Li. 2013. "Evidence on the Impact of Sustained Exposure to Air Pollution on Life Expectancy from China's Huai River Policy." *Proceedings of the National Academy of Sciences* 110(32): 12936–41.

Davis, Lucas W. 2012. "Prospects for Nuclear Power." *Journal of Economic Perspectives* 26(1): 49–66.

Deschênes, Olivier, and Michael Greenstone. 2011. "Climate Change, Mortality, and Adaptation: Evidence from Annual Fluctuations in Weather in the US." *American Economic Journal: Applied Economics* 3(4): 152–85.

Deutch, John M., Charles W. Forsberg, Andrew C. Kadak, Mujid S. Kazimi, Ernest J. Moniz, and John E. Parsons. 2009. *Update of the MIT 2003 Future of Nuclear Power.* An Interdisciplinary MIT Study. http://web.mit.edu/nuclearpower/pdf/nuclearpower-update2009.pdf.

Dyni, John R. 2006. "Geology and Resources of Some World Oil-Shale Deposits." USGS Scientific Investigations Report 2005-5294. http://pubs.usgs.gov/sir/2005/5294/pdf/sir5294_508.pdf.

Graff Zivin, Joshua S., Matthew J. Kotchen, and Erin T. Mansur. 2014. "Spatial and Temporal Heterogeneity of Marginal Emissions: Implications for Electric Cars and Other Electricity-Shifting Policies." *Journal of Economic Behavior and Organization* 107(Part A): 248–68.

Greenstone, Michael, Elizabeth Kopits, and Ann Wolverton. 2013. "Developing a Social Cost of Carbon for US Regulatory Analysis: A Methodology and Interpretation." *Review of Environmental Economics and Policy* 7(1): 23–46.

Hubbert, M. King. 1956. "Nuclear Energy and the Fossil Fuels." Publication no. 95. Shell Development Company, Exploration Production and Research Division. http://www.hubbertpeak.com/hubbert/1956/1956.pdf.

Intergovernmental Panel on Climate Change (IPCC). 2013. "Summary for Policymakers." In: *Climate Change 2013: The Physical Science Basis,* edited by T. F. Stocker et al. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

International Energy Agency. 2015. *World Energy Outlook 2015.* http://www.worldenergyoutlook.org/weo2015/.

Knittel, Christopher R. 2011. "Automobiles on Steroids: Product Attribute Trade-offs and Technological Progress in the Automobile Sector." *American Economic Review* 101(7): 3368–99.

Knittel, Christopher R., Douglas L. Miller, and Nicholas J. Sanders. Forthcoming. "Caution, Drivers! Children Present: Traffic, Pollution, and Infant Health." *Review of Economics and Statistics.*

McGlade, Christophe, and Paul Ekins. 2015. "The Geographical Distribution of Fossil Fuels Unused When Limiting Global Warming to 2°C." *Nature* 517(January 8): 187–90.

McKinsey & Company. 2009. *Pathways to a Low-Carbon Economy.* http://www.mckinsey.com/client_service/sustainability/latest_thinking/pathways_to_a_low_carbon_economy.

Mills Andrew, and Ryan Wiser. 2010. "Implications of Wide-Area Geographic Diversity for Short-Term Variability of Solar Power." LBNL-3884E, Lawrence Berkeley National Laboratory, Berkeley, California.

Murray, James, and David King. 2012. "Climate Policy: Oil's Tipping Point Has Passed." *Nature* 481(January 26): 433–35.

Nykvist, Björn and Måns Nilsson. 2015. "Rapidly Falling Costs of Battery Packs for Electric Vehicles." *Nature Climate Change* 5: 329–332.

Peterson, Scott B., and Jeremy J. Michalek. 2013. "Cost-Effectiveness of Plug-in Hybrid Electric Vehicle Battery Capacity and Charging Infrastructure Investment for Reducing US Gasoline Consumption." *Energy Policy* 52: 429–38.

Rutledge, David B. 2013. "Projections for Ultimate World Coal Production from Production Histories Through 2012." Presentation to the 125th Anniversary Annual Meeting and Expo of the Geological Society of America, Annual Meeting, Denver, CO, October 2013.

Shellberger, Michael, Ted Nordhaus, Alex Trembath, and Jesse Jenkins. 2012. "Where the Shale Gas Revolution Came From." Technical Report, The Breakthrough Institute, May.

US Department of Energy. 2014. "EV Everywhere Grand Challenge." DOE/EE-1024. http://energy.gov/sites/prod/files/2014/02/f8/eveverywhere_road_to_success.pdf.

US Energy Information Administration (EIA). 2012. "Potential of Gas Hydrates is Great, But Practical Development is Far Off." November 7. http://www.eia.gov/todayinenergy/detail.cfm?id=8690.

US Energy Information Administration (EIA). 2013. *Technically Recoverable Shale Oil and Shale Gas Resources: An Assessment of 137 Shale Formations in 41 Countries Outside the United States.* June. http://www.eia.gov/analysis/studies/worldshalegas/pdf/fullreport.pdf.

US Energy Information Administration (EIA). No date. "International Energy Statistics." Database. http://www.eia.gov/cfapps/ipdbproject/IEDIndex3.cfm.

US Environmental Protection Agency. No date. Global Greenhouse Gas Emissions Data. http://www3.epa.gov/climatechange/ghgemissions/global.html.

Wolak, Frank. 2015. "Mean versus Standard Deviation Trade-offs in Wind and Solar Energy Investments: The Case of California." Unpublished paper, Stanford University.

Journal of Economic Perspectives—Volume 30, Number 1—Winter 2016—Pages 139–160

Forty Years of Oil Price Fluctuations: Why the Price of Oil May Still Surprise Us

Christiane Baumeister and Lutz Kilian

It has been 40 years since the oil crisis in 1973/74, which also coincided with the emergence of a new regime in the global market for crude oil, in which oil prices have been largely free to fluctuate in response to the forces of supply and demand (Dvir and Rogoff 2010; Alquist, Kilian, and Vigfusson 2013). The crisis arose when the price of imported oil nearly quadrupled over the course of a quarter, forcing substantial adjustments in oil-consuming countries. To make matters worse, some governments in industrialized countries responded by imposing ceilings on the price of domestically produced crude oil and on the price of refined oil products such as gasoline, causing gasoline shortages and long lines at gas stations. In addition, many governments introduced speed limits, banned automobile traffic on Sundays, or limited retail gasoline purchases (for example, Ramey and Vine 2011). Hence, pictures of long lines at gas stations and empty highways have shaped the collective memory of the 1973/74 oil crisis, even though in reality neither phenomenon was an inevitable consequence of the underlying rise in the price of crude oil.

Although sharp oil price increases had occurred at irregular intervals throughout the post–World War II period, as documented in Hamilton (1983, 1985), none of

■ *Christiane Baumeister is Assistant Professor of Economics at the University of Notre Dame, Notre Dame, Indiana, and Research Affiliate at the Centre of Economic Policy Research, London, United Kingdom. Lutz Kilian is Professor of Economics at the University of Michigan, Ann Arbor, Michigan, and Research Fellow at the Centre of Economic Policy Research, London, United Kingdom. Their email addresses are cjsbaumeister@gmail.com and lkilian@umich.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.1.139

doi=10.1257/jep.30.1.139

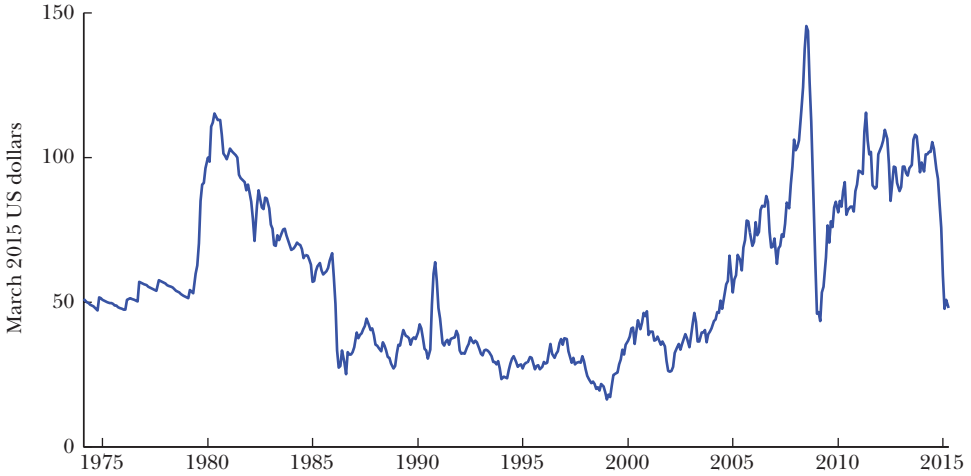
these increases was comparable in magnitude to the increase in the price of oil in the last quarter of 1973. In fact, prior to 1973, the US price of oil had been regulated by government agencies, resulting in extended periods of constant oil prices, interrupted only by infrequent adjustments, which tended to coincide with exogenous oil supply disruptions in the Middle East. This policy resulted in occasional sharp spikes in the growth rate of the inflation-adjusted price of crude oil.

The US system of oil price regulation came to an end starting in the early 1970s, when the United States no longer had any spare capacity in domestic oil production to satisfy its growing domestic demand for oil and became increasingly dependent on oil imports from the Middle East, the price of which could not be regulated domestically (Yergin 1992). When the price of imported crude oil quadrupled in 1973/74, imposing lower ceilings on the price of domestically produced crude oil soon proved impractical. The price of oil as measured per barrel of the West Texas Intermediate (WTI) benchmark—a particular grade of light and sweet crude oil commonly traded in the United States—rose from $4.31 per barrel in September 1973 to $10.11 in January 1974. Although the last vestiges of the regulation of the price of domestic crude oil in the United States persisted until the early 1980s, for all practical purposes, there was a structural break in the time series process governing the WTI price of crude oil in early 1974, with the real price of oil fluctuating in response to supply and demand shocks much like other real industrial commodity prices. It is this modern era of oil markets that our discussion focuses on.

Figure 1 plots the real price of oil (expressed in March 2015 dollars) starting in January 1974. It shows substantial fluctuations in the real price of oil in recent decades with no obvious long-run trend. The literature has identified a number of potential determinants of oil price fluctuations, including: 1) shocks to global crude oil production arising from political events in oil-producing countries, the discovery of new fields, and improvements in the technology of extracting crude oil; 2) shocks to the demand for crude oil associated with unexpected changes in the global business cycle; and 3) shocks to the demand for above-ground oil inventories, reflecting shifts in expectations about future shortfalls of supply relative to demand in the global oil market.

In this article, we review the causes of the major oil price fluctuations since 1973/74, episode by episode. Although economists have made great strides in recent years in understanding the oil price fluctuations in Figure 1 with the benefit of hindsight, some of the variation in the price of oil over the last 40 years was clearly unexpected at the time. We discuss alternative measures of oil price expectations employed by central banks, by economists, and by households as well as measures of financial market expectations of the price of oil. Although some oil price expectations measures can be shown to be systematically more accurate than others, all oil price expectations are subject to error. The reason is that even if we understand the determinants of the price of oil, predicting these determinants can be very difficult in practice. We discuss in the context of concrete examples why it is so difficult to predict the determinants of the price of oil.

Figure 1
Inflation-Adjusted WTI Price of Crude Oil, 1974.1–2015.3



Source: US Energy Information Administration.
Note: The West Texas Intermediate (WTI) oil price series has been deflated with the seasonally adjusted US consumer price index for all urban consumers.

The gap between the price of oil that was expected and its eventual outcome represents an oil price "shock." Such surprise changes in the price of oil have been considered important in modeling macroeconomic outcomes in particular. We demonstrate how much the timing and magnitude of oil price shocks may change with the definition of the oil price expectations measure. We make the case that the oil price expectations measure required for understanding economic decisions need not be the most accurate measure in a statistical sense, and we illustrate that the same change in oil prices may be perceived quite differently by households, policymakers, financial markets, and economists, depending on how they form expectations. This insight has potentially important implications for understanding and modeling the transmission of oil price shocks.

Historical Episodes of Major Fluctuations in the Real Price of Oil

The literature on the causes of oil price fluctuations has evolved substantially since the early 1980s. Initially, all major oil price fluctuations were thought to reflect disruptions of the flow of global oil production associated with exogenous political events such as wars and revolutions in OPEC member countries (for example, Hamilton 2003).[1] Subsequent research has shown that this explanation is only one

[1] OPEC refers to Organization of Petroleum Exporting Countries, which was founded in 1960.

among many, and not as important as originally thought. In fact, most major oil price fluctuations dating back to 1973 are largely explained by shifts in the demand for crude oil (for example, Barsky and Kilian 2002, 2004; Kilian 2009a; Kilian and Murphy 2012, 2014; Bodenstein, Guerrieri, and Kilian 2012; Lippi and Nobili 2012; Baumeister and Peersman 2013; Kilian and Hicks 2013; Kilian and Lee 2014).[2] By far the most important determinant of the demand for oil has been shifts in the flow (or consumption) demand for oil associated with the global business cycle. As the global economy expands, so does demand for industrial raw materials including crude oil, putting upward pressure on the price of oil. At times there also have been important shifts in the demand for stocks (or inventories) of crude oil, reflecting changes to oil price expectations. Such purchases are not made because the oil is needed immediately in the production of refined products such as gasoline or heating oil, but to guard against future shortages in the oil market. Historically, inventory demand has been high in times of geopolitical tension in the Middle East, low spare capacity in oil production, and strong expected global economic growth.

The 1973/74 Oil Crisis

At first sight, the oil price shock of 1973/74 has the appearance of a negative shock to the supply of crude oil in that the quantity of crude oil produced fell in the last quarter of 1973 and the price of oil increased, consistent with a shift of the supply curve to the left along the demand curve. Indeed, this is the traditional explanation for this oil price increase advanced by Hamilton (2003). It is common to refer to the war between Israel and a coalition of Arab countries that took place between October 6 and 26, 1973, as the cause of this supply shock. This explanation may conjure up images of burning oil fields, but actually there was no fighting in any of the Arab oil-producing countries in 1973 and no oil production facilities were destroyed. Instead, this war took place in Israel, Egypt, and Syria. None of these countries was a major oil producer or a member of OPEC, for that matter. Thus, the disruption of the flow of oil production that took place in the last quarter of 1973 was not a direct effect of the war. Rather, Arab OPEC countries deliberately cut their oil production by 5 percent starting on October 16, 1973, ten days into the Arab–Israeli War, while raising the posted price of their oil, followed by the announcement of an additional 25 percent production cut on November 5, ten days after the war had ended.

Hamilton (2003) attributes the Arab oil production cuts in October and November 1973 entirely to the Arab oil embargo against selected Western countries, which lasted from October 1973 to March 1974, interpreting this oil embargo as an extension of the military conflict by other means rather than an endogenous response to economic conditions. There is, however, an alternative interpretation of the same data that does not rely on the war as an explanation. Barsky and

[2] In related work, Carter, Rausser, and Smith (2011) conclude that similar results hold for commodity prices more generally, noting that commodity price booms over the last four decades have been preceded by unusually high economic growth.

Kilian (2002) draw attention to the fact that in early 1973 the price of crude oil received by Middle Eastern oil producers was effectively fixed as a result of the 1971 Tehran/Tripoli agreements between oil companies and governments of oil producing countries in the Middle East. These five-year agreements set the price of oil received by the host government for each barrel of oil extracted in exchange for assurances that the government would allow foreign oil companies to extract as much oil as they saw fit (Seymour 1980, p. 80). When global demand for oil accelerated in 1972–73, reflecting a worldwide economic boom, many Middle Eastern countries were operating close to capacity already and unable to increase oil output; whereas others, notably Saudi Arabia and Kuwait, had the spare capacity to increase their output, and allowed their oil production to be increased, albeit reluctantly. This reluctance can be attributed to the fact that the posted price agreed upon in 1971 might have been reasonable at the time, but was quickly eroded in real terms as a result of a depreciating US dollar and rising US inflation. This development caused increasing Arab opposition to the Tehran/Tripoli agreements that intensified in March of 1973 and culminated in the repudiation of the agreements on October 10, 1973, with oil producers deciding to produce less oil at higher prices.

This reaction makes economic sense even in the absence of any monopoly power by oil producers. Under this interpretation, a substantial fraction of the observed decline in Arab oil output in late 1973 was simply a reversal of an unusual increase in Saudi and Kuwaiti oil production that had occurred earlier that year in fulfillment of the Tehran/Tripoli agreements. Moreover, the decision to reduce oil production and the objective of raising the oil price was clearly motivated by the cumulative effects of the dollar devaluation, unanticipated US inflation, and high demand for oil fueled by strong economic growth, making this oil price increase endogenous with respect to global macroeconomic conditions.[3]

There is, of course, no reason to expect the price of oil charged by Arab oil producers as of January 1974 to be the equilibrium price necessarily. This price was set on the basis of negotiations among oil producers, not by the market. There is evidence, however, that the negotiated price was in fact close to the equilibrium value. A good indication of the shadow price of crude oil is provided by the steady increase in commonly used indices of non-oil industrial commodity prices between November 1971 and February 1974. In the absence of contractual constraints, one would have expected the price of oil to grow at a similar rate in response to increased global demand. Kilian (2009b), shows that non-oil industrial commodity prices over this period increased by 75 percent as much as the price of crude oil (with some individual commodity prices quadrupling, not unlike the price of oil), suggesting that at most 25 percent of the oil price increase of 1973/74 was caused

[3] A detailed analysis in Kilian (2008a) shows that the observed changes in the price of oil and in the quantity of oil produced in the Middle East over the course of the year 1973 is consistent with this interpretation. Notably, only Kuwait and Saudi Arabia reduced their oil output in October 1973 and only to the extent required to return to normal levels of oil production, suggesting that the war was immaterial as a motive for the October production cut.

by exogenous oil supply shocks. This evidence suggests that much of the oil crisis of 1973/74 actually was driven by increased demand for oil rather than reductions in oil supply. This conclusion is also consistent with the predictions of regressions of changes in the price of oil on direct measures of exogenous OPEC oil supply shocks (see Kilian 2008a). These regressions suggest that it is difficult to explain more than 25 percent of the 1973 oil price increase based on exogenous OPEC supply shocks.

The 1979/80 Oil Crises

The oil crisis of 1973/74 was followed by a second major oil crisis in 1979/80, when the price of West Texas Intermediate crude oil rose from less than $15 per barrel in September 1978 to almost $40 in April 1980. As in 1973/74, governments responded to rising oil prices by rationing gasoline and enforcing price controls, causing the recurrence of long lines at gas stations. The traditional view, expressed in Hamilton (2003), has been that this surge in the price of oil was caused by the reduction in Iranian oil production following the Iranian Revolution. As noted in Barsky and Kilian (2002), the timing of events casts doubt on this interpretation. The Iranian Revolution started gradually in late 1978, culminating in the departure of the Shah of Iran in January 1979 and the arrival of Ayatollah Khomeini in February 1979. The biggest Iranian production shortfalls occurred in January and February 1979. Iranian oil production started recovering in March. Given increased oil production in Saudi Arabia in direct response to the Iranian Revolution, the shortfall of OPEC oil output in January 1979 was 8 percent relative to September 1978. By April, the shortfall of OPEC output was zero percent. The price of oil did not increase substantially before May 1979. Even in April 1979, the WTI oil price was still under $16 per barrel (up about $1 from $14.85 prior to the Iranian Revolution), yet within a year the WTI price would reach a peak level of $40 per barrel in April 1980. The same pattern is also found in oil price series not regulated by the government, such as the US refiners' acquisition cost of crude oil imports. It is not clear why the effect of an oil supply shock on the price of oil would be delayed for so long. Thus, the timing of this oil supply shock makes it an unlikely candidate for explaining the 1979 oil price increase.

As stressed in Kilian and Murphy (2014), this does not mean that the Iranian Revolution did not matter for the price of oil, but that it mattered because it affected oil price expectations rather than because it affected the flow of oil production. Empirical oil market models that allow for both oil demand and oil supply shocks to affect the price of oil confirm that oil supply shocks played a minor role for the 1979 oil price increase, but suggest that about one-third of the cumulative price increase was associated with increased inventory demand in anticipation of future oil shortages, presumably reflecting geopolitical tensions between the United States and Iran and between Iran and its neighbors, but also expectations of high future demand for oil from a booming global economy. This evidence of rising inventory demand starting in May 1979 is also consistent with anecdotal evidence from oil market participants (for example, Yergin 1992). The remaining two-thirds of the cumulative oil price increase in 1979 are explained by the cumulative effects of flow demand shocks

triggered by an unexpectedly strong global economy, not unlike during the first oil crisis (for example, Kilian 2009a; Kilian and Murphy 2014).

The 1980s and 1990s

Hamilton (2003) also identifies a large exogenous oil supply disruption associated with the outbreak of the Iran–Iraq War, which lasted from 1980 until 1988. In late September 1980, Iraq invaded Iran, causing the destruction of Iranian oil facilities and disrupting oil exports from both Iran and Iraq. This event was followed by an increase in the WTI price of oil from $36 per barrel in September to $38 in January 1981, which by all accounts must be attributed to this oil supply disruption. This episode is instructive because it represents an example of an oil supply shock occurring in the absence of major shifts in oil demand. There is little evidence of this shock triggering a large price response, consistent with more formal estimates from structural oil market models.

The early 1980s saw a systematic decline in the price of oil from its peak in April 1980. One reason was the shift in global monetary policy regimes toward a more contractionary stance, led by Paul Volcker's decision to raise US interest rates. The resulting global recession lowered the demand for oil and hence the price of oil. This decline was further amplified by efforts to reduce the use of oil in industrialized countries. In addition, declining prospects of future economic growth in conjunction with higher interest rates made it less attractive to hold stocks of oil, causing a sell-off of the oil inventories accumulated in 1979. Finally, one of the legacies of the first oil crisis had been that numerous non-OPEC countries including Mexico, Norway, and the United Kingdom responded to persistently high oil prices by becoming oil producers themselves or by expanding their existing oil production. Given the considerable lag between exploration and production, it was only in the early 1980s, that this supply response to earlier oil price increases became quantitatively important. OPEC's global market share fell from 53 percent in 1973 to 43 percent in 1980 and 28 percent in 1985. The increase in non-OPEC oil production put further downward pressure on the price of oil.

OPEC attempted to counteract the decline in the price of oil in the early 1980s. Indeed, this is the first time in its history (and the only time) that OPEC took a proactive role in trying to influence the price of oil (also see Almoguera, Douglas, and Herrera 2011).[4] When OPEC agreements to jointly restrict oil production in an effort to prop up the price of oil proved ineffective, with many OPEC members cheating on OPEC agreements, as predicted by the economic theory of cartels (for example, Green and Porter 1984), Saudi Arabia decided to stabilize the price of oil on its own by reducing Saudi oil production. Skeet (1988) offers

[4] The literature has not been kind to the view that OPEC since 1973 has acted as a cartel that sets the price of oil or that controls the price of oil by coordinating oil production among OPEC members. For example, Alhajji and Huettner (2000) stress that OPEC does not fit the theory of cartels. Smith (2005) finds that there is no conclusive evidence of OPEC acting as a cartel. Cairns and Calfucura (2012) conclude that OPEC has never been a functioning cartel. For further discussion also see Almoguera, Douglas, and Herrera (2011) and Colgan (2014).

a detailed discussion of how these policies were implemented. As Figure 1 shows, this approach did not succeed. The price of oil continued to fall in the early 1980s, albeit at a slower pace. The resulting losses in Saudi oil revenue proved so large that by the end of 1985, Saudi Arabia was forced to reverse its policy of restricting oil production. The result was a sharp fall in the price of oil in 1986, caused not only by the resumption of Saudi oil production, but more importantly by a reduction in inventory demand for oil, given that OPEC had shown itself to be unable to sustain a higher price of oil (Kilian and Murphy 2014).

Given the abundance of crude oil supplies in the world relative to oil demand, the ongoing Iran–Iraq War had little effect on the price of oil in the 1980s, notwithstanding considerable damage to oil shipping in the Persian Gulf with as many as 30 attacks on oil tankers in a given month. It took the invasion of Kuwait in August of 1990, followed by the Persian Gulf War, which was directed at ejecting Saddam Hussein from Kuwait, to generate a sharp increase in the price of oil. The disruption in Iraqi and Kuwaiti oil production associated with this war played an important role in causing this spike in the price of oil, but an equally important determinant was higher demand for oil inventories in anticipation of a possible attack on Saudi oil fields (Kilian and Murphy 2014). Only in late 1990, when the coalition led by the United States had moved enough troops to Saudi Arabia to forestall an invasion of Saudi Arabia, these fears subsided and the price of oil fell sharply along with inventory demand. Without this expectational element, it would be difficult to explain the quick return to lower oil prices in 1991, given that oil production from Kuwait and Iraq was slow to recover (Kilian 2008a).

In the late 1990s, the price of oil weakened further. By December 1998, the WTI price of oil reached an all-time low in recent history of $11, when only two years earlier oil had been trading at $25. This slide was largely associated with reduced demand for crude oil, arguably caused by the Asian financial crisis of mid-1997, which in turn was followed by economic crises in other countries including Russia, Brazil, and Argentina. The recovery in the price of oil starting in 1999 reflected a combination of factors including higher demand for oil from a recovering global economy, some cuts in oil production, and increased inventory demand in anticipation of tightening oil markets (Kilian and Murphy 2014).

This recovery was followed by two major exogenous oil supply disruptions in late 2002 and early 2003 that in combination rivaled the magnitude of the oil supply disruptions of the 1970s (Kilian 2008a). One was a sharp drop in Venezuelan oil production caused by civil unrest in Venezuela; the other was the disruption of oil production associated with the 2003 Iraq War. The production shortfalls in Iraq and Venezuela were largely offset by increased oil production elsewhere, however. Moreover, compared with 1990, there was less concern that this war would affect oil fields in Saudi Arabia, especially after the US-led ground offensive proved successful, with missile attacks being the main threat at the beginning of the war. As a result, there was only a modest shift in inventory demand. Indeed, this oil supply shock episode is remarkable mainly because the price of oil proved resilient to geopolitical events. The price of oil only briefly spiked, adding approximately an extra $6 per barrel.

From the Great Surge of 2003–08 to the Global Financial Crisis

The most remarkable surge in the price of oil since 1979 occurred between mid-2003 and mid-2008 with the WTI price climbing from $28 to $134 per barrel. There is widespread agreement that this price surge was not caused by oil supply disruptions, but by a series of individually small increases in the demand for crude oil over the course of several years. Kilian (2008b), Hamilton (2009), and Kilian and Hicks (2013), among others, have made the case that these demand shifts were associated with an unexpected expansion of the global economy and driven by strong additional demand for oil from emerging Asia in particular. Because oil producers were unable to satisfy this additional demand, the price of oil had to increase. This view is consistent with estimates from empirical models of the global oil market, which attribute the bulk of the cumulative increase in the price of oil to flow demand shocks (Kilian 2009a; Baumeister and Peersman 2013; Kilian and Murphy 2014). Only in the first months of 2008 is there any evidence of increased inventory demand (Kilian and Lee 2014).

An alternative view among some observers has been that this surge in the price of oil in the physical market is unprecedented and can only be explained as the result of speculative positions taken by financial traders in the oil futures market. This literature has been reviewed in depth in Fattouh, Killian, and Mahadeva (2013). There is no persuasive evidence in support of this financial speculation hypothesis, which in fact is at odds with estimates of standard economic models of markets for storable commodities (for example, Alquist and Kilian 2010; Kilian and Murphy 2014; Knittel and Pindyck forthcoming).

The financial crisis of 2008 illustrates the powerful effects of a sharp drop in the demand for industrial commodities on the price of these commodities. As orders for industrial commodities worldwide were sharply curtailed in the second half of 2008 in anticipation of a major global recession, if not depression, the demand for commodities such as crude oil plummeted, causing a fall in the price of oil from $134 per barrel in June 2008 to $39 in February 2009. It is noteworthy that shifts in the demand for industrial commodities such as crude oil may have much larger amplitude than the corresponding changes in global real GDP because global real GDP consists to a large extent of consumption, which remained much more stable during the crisis. When it became clear in 2009 that the collapse of the global financial system was not imminent, the demand for oil recovered to levels prevailing in 2007, and the price of oil stabilized near $100 per barrel.

There have been a number of smaller demand and supply shocks in the oil market between 2010 and early 2014. For example, events such as the Libyan uprising in 2011 were associated with an increase in the price of oil. Kilian and Lee (2014) estimate that the Libyan crisis caused an oil price increase of somewhere between $3 and $13 per barrel, depending on the model specification. Likewise, tensions with Iran in 2012 account for an increase of between $0 and $9 per barrel. An additional development since 2011 has been a widening of the spread between the two main benchmark prices for oil, the West Texas Intermediate and Brent prices, with WTI oil trading at a discount, reflecting a local glut of light sweet crude oil in the central United States driven by increased US shale oil production (Kilian

2014a). As a result, the WTI price of crude oil is no longer representative for the price of oil in global markets, and it has become common to use the price of Brent crude oil as a proxy for the world price in recent years.

Following a long period of relative price stability, between June 2014 and January 2015 the Brent price of oil fell from $112 to $47 per barrel, providing yet another example of a sharp decline in the price of oil, not unlike those in 1986 and 2008. In Baumeister and Kilian (2015a), we provide the first quantitative analysis of the $49 per barrel drop in the Brent price between June and December 2014. We conclude that about $11 of this decline was associated with a decline in global real economic activity that was predictable as of June 2014 and reflected in other industrial commodity prices as well. An additional decline in the Brent price of $16 was predictable as of June 2014 on the basis of shocks to actual and expected oil production that took place prior to July 2014. These shocks likely reflected the unexpected growth of US shale oil production but also increased oil production in other countries including Canada and Russia. The remaining decline of $22 in the Brent price is explained by two shocks taking place in the second half of 2014. One is a $9 decline explained by a shock to the storage demand for oil in July 2014; a further $13 decline is explained by an unexpected weakening of the global economy in December 2014.

How to Measure Oil Price Expectations

Although economists have made great strides in recent years in understanding historical oil price fluctuations, as illustrated in the preceding section, some of the variation in the price of oil over the last 40 years was unexpected at the time. The extent to which oil price fluctuations are unexpected depends on how expectations are formed. Below we introduce four alternative measures of oil price expectations that may be viewed as representative for the oil price expectations of economists, policymakers, financial market participants, and consumers, respectively.

Economists' Oil Price Expectations

One common approach to constructing oil price expectations is to relate the price of oil to its own past values as well as past values of other key determinants of the price of oil suggested by economic theory (for example, Alquist, Kilian, and Vigfusson 2013). This is the central idea underlying vector autoregression (VAR) models of the global oil market (Kilian 2008b, 2009a; Baumeister and Peersman 2013). In our empirical analysis, we employ a VAR model specification in the tradition of Kilian and Murphy (2014) that includes the real price of oil, global crude oil production, global real economic activity, and changes in global crude oil stocks. We refer to the implied expectation of the price of oil as the economists' expectation.

Policymakers' Oil Price Expectations

Of course, oil price expectations based on regression models need not be representative for the views held by financial market participants or by firms and

households in the economy. One possibility is that financial market participants have more information than can be captured by econometric models. It is equally possible for financial market participants to ignore, misinterpret, or miss information captured by model-based oil price predictions, especially if that information is costly to obtain. A natural source of information about the market expectation of the price of oil is the price of oil futures contracts. Futures contracts are financial instruments that allow traders to lock in today a price at which to buy a fixed quantity of crude oil at a predetermined date in the future. The most common approach to inferring the expected price of oil for immediate delivery in the physical market (also known as the spot price) has been to treat the price of the oil futures contract of maturity h as the h-period ahead market expectation of the nominal price of crude oil. If today's price of a futures contract expiring a year from now is $50, for example, then the expectation of the spot price of oil one year from now is $50. This is how the International Monetary Fund forms oil price expectations, and this has been common practice at many central banks in the world, as discussed in Alquist, Kilian, and Vigfusson (2013), which is why we refer to this approach as the policymakers' oil price expectation.

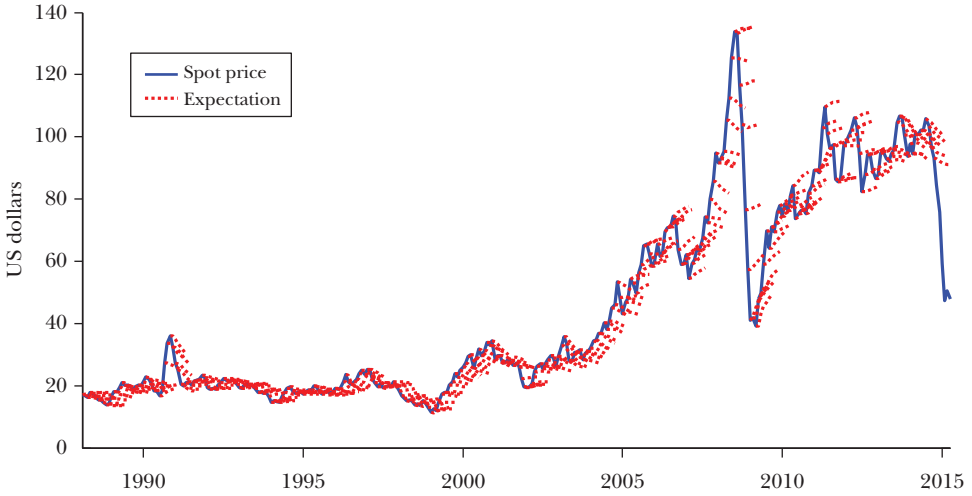
Financial Market Oil Price Expectations

The use of futures prices as measures of market expectations, however, is valid only if the risk premium—defined as the compensation arbitrageurs receive for assuming the price risk faced by hedgers in the oil futures market—is negligible. This assumption is questionable. Hamilton and Wu (2014) document that there is a large horizon-specific time-varying risk premium in the oil futures market. This risk premium varies with the hedging demands of oil producers and refiners and the willingness of financial investors to take the other side of hedging contracts. The oil price expectation may be recovered by subtracting the Hamilton–Wu estimates of the risk premium from the oil futures price for a given horizon. In a comparison of alternative risk premium estimates, this specification has been shown to produce the most reliable oil price expectations measure overall (Baumeister and Kilian 2015a). We refer to the latter expectation as the financial market expectation of the price of oil.

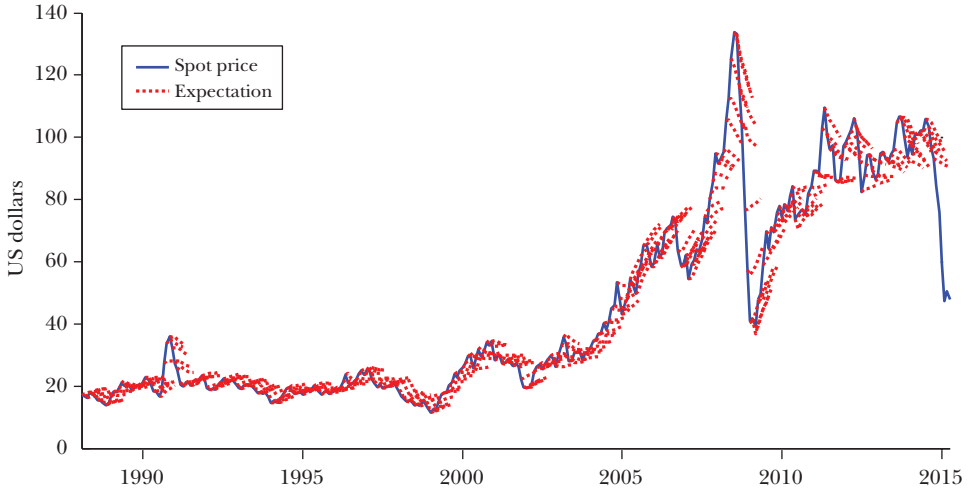
The policymaker's expectation based on the oil futures price and the financial market expectation proposed in Baumeister and Kilian (2015a) differ not only quantitatively, but also qualitatively. For expository purposes, our discussion focuses on expectations of the West Texas Intermediate price of crude oil; analogous results could also be constructed for the Brent price. We consider horizons of up to six months. The sample period starts in January 1988 and extends to the end of 2014. Figure 2A treats the current oil futures price with a maturity of h months as the expectation for the spot price of oil in h months. In contrast, Figure 2B plots the corresponding financial market expectation of the price of oil obtained by subtracting from the current WTI futures price of maturity h the risk premium estimate for horizon h implied by the term structure model of Hamilton and Wu (2014).

Figure 2
Alternative Expectations Measures Based on WTI Futures Prices

A: Monthly Oil Price Expectations Measure Obtained from the Oil Futures Curve



B: Monthly Financial Market Oil Price Expectations Obtained from the Risk-Adjusted Futures Curve



Notes: In Figure 2A, the expectation of the West Texas Intermediate (WTI) spot price for a horizon of h months is measured by the WTI oil futures price at maturity h. This approach ignores the possible presence of a risk premium (Baumeister and Kilian 2015a). In Figure 2B, the expectation of the WTI spot price for a horizon of h months is measured by subtracting from the WTI oil futures price at maturity h the estimated risk premium for that horizon. All risk premium estimates are constructed from the weekly term structure model proposed by Hamilton and Wu (2014), as implemented in Baumeister and Kilian (2015a).

Figures 2A and 2B illustrate that adjusting the futures price for the risk premium may matter a lot in measuring oil price expectations. For example, at the peak of the oil price in mid-2008 and in the subsequent months the term structure of futures prices in Figure 2A slopes upwards, seemingly implying expectations of rising oil prices, whereas the risk-adjusted futures prices in Figure 2B indicate expectations of sharply falling oil prices. Only after the spot price of crude oil fell below about $85 in late 2008, did participants in the futures market expect the price of oil to recover. Similar patterns can also be found around the smaller oil price peaks of 2011 and 2012.

Another important difference is that during the long surge in the spot price between 2003 and early 2008, the futures curve in Figure 2A mostly suggests expectations of falling oil prices, whereas the risk-adjusted futures curve in Figure 2B often indicates much more plausible expectations of rising oil prices. Only when the spot price surpassed about $100 per barrel, did the risk-adjusted futures price become more bearish than the unadjusted futures price. Finally, following the invasion of Kuwait in 1990, the unadjusted futures curve in Figure 2A suggests expectations of much more rapidly falling oil prices than the risk-adjusted futures curve in Figure 2B.

Consumers' Oil Price Expectations

Yet another approach to measuring oil price expectations is to focus more directly on the expectations of firms or households. There are no survey data on the oil price expectations of US households or US manufacturing firms (or of oil companies, for that matter), but recent research by Anderson, Kellogg, Sallee, and Curtin (2011) has shown that households in the Michigan Survey of Consumers typically form expectations about the real (or inflation-adjusted) price of gasoline according to a simple no-change model such that the nominal gasoline price is expected to grow at the rate of inflation. Given that the price of gasoline is primarily determined by the price of crude oil, a reasonable conclusion is that consumers forecast the real and nominal prices of crude oil along much the same lines, allowing us to proxy consumer expectations about the nominal price of oil based on the current price of oil and an inflation forecast. Because the inflation expectations data in the Michigan Survey of Consumers is limited to selected horizons, in practice, we rely on the fixed coefficient gap model inflation forecast proposed in Faust and Wright (2013). Given that the inflation component in oil price forecasts is small at short horizons, as shown in Alquist, Kilian, and Vigfusson (2013), the difference is likely to be negligible.

The use of such simple prediction rules makes perfect economic sense when consumers do not have access to more sophisticated oil price forecasts. After all, consumers cannot be expected to have the time and resources to become experts in oil price forecasting or to make sense of the range of competing oil price forecasts produced by experts. Indeed, a good case can be made that in thinking about and in modeling households' purchase decisions, it is households' own oil price expectations that matter even if those expectations are not as accurate as some

alternative model-based forms of oil price expectations. We refer to this simple rule of thumb as the consumer oil price expectation.

What Is an Oil Price Shock?

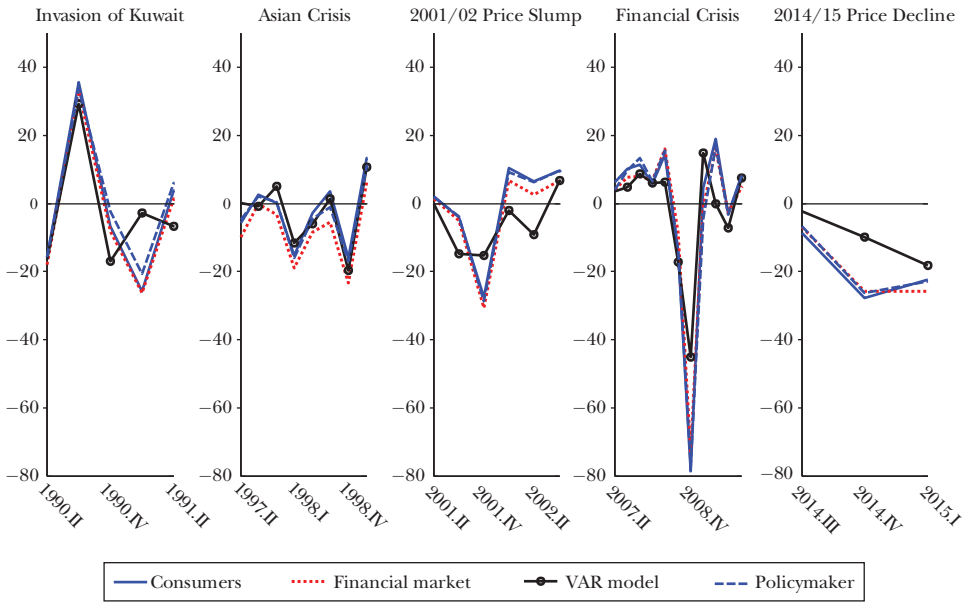
The unanticipated or surprise component of a change in the price of oil is referred to as an oil price shock. By comparing oil price expectations to subsequent outcomes, we may obtain a direct measure of the magnitude of the oil price shock. Clearly, whether an oil price shock occurred, and how large this shock was, depends on which measure of the oil price expectations we use. This question has not received much attention to date. Below we compare: 1) the oil price shocks perceived by financial markets, based on the oil price expectations shown in Figure 2B; 2) the oil price shocks perceived by policymakers, as shown in Figure 2A; 3) oil price shocks perceived by consumers employing a simple rule of thumb for predicting oil prices; and 4) oil price shocks as measured by economists employing oil market vector autoregressive (VAR) models. For expository purposes, we again focus on the WTI price of oil.

Figure 3 shows all four measures of oil price shocks of interest for five specific episodes, starting in 1988.II, namely the invasion of Kuwait in 1990, the Asian financial crisis of 1997, the 2001/02 oil price slump, the financial crisis in 2008, and the 2014/15 oil price decline. In each of these episodes, the oil price shocks implied by the vector autoregressive (VAR) model are smaller on average than those implied by any of the other three expectations measures, but the differences are most pronounced in the last two episodes. For example, following the global financial crisis, in the last quarter of 2008, the consumer oil price shock was −79 percent, and the policymaker oil price shock was −77 percent, compared with only −73 percent based on financial market expectations, suggesting that financial markets were able to anticipate the oil price decline in the fourth quarter somewhat more accurately than consumers and policymakers. Even more striking is that the VAR oil market model predicted a much larger decline in oil prices than financial markets did, resulting in an oil price shock in the fourth quarter of only −45 percent.

Similarly, neither consumers nor policymakers nor financial markets anticipated the 2014/15 oil price drop, which was reflected in large negative oil price shocks of between −7 and −9 percent in the third quarter of 2014, between −26 and −28 percent in the fourth quarter of 2014, and between −22 and −26 percent in the first quarter of 2015. Again, this decline was predicted to a considerable extent by the vector autoregressive (VAR) model, which implies much smaller model-based oil price shocks of −2 percent for 2014.III, −10 percent for 2014.IV, and of −18 percent for 2015.I.

Figure 3 illustrates that it can make a difference whether we take the consumer's perspective, the policymaker's perspective, the financial market perspective, or the economist's perspective in measuring oil price shocks. The extent of the differences differs by episode. Overall, oil price shocks are largest from the

Figure 3
Quarterly Shocks to Nominal WTI Price of Oil by Episode
(percent)



Notes: Each oil price shock series is constructed by averaging the monthly oil price expectations by quarter and expressing this average as a percent deviation from the quarterly average of the monthly oil price outcomes. The policymakers' expectation corresponds to the unadjusted West Texas Intermediate (WTI) oil futures price. The financial market expectation is constructed by subtracting the Hamilton and Wu (2014) risk premium estimate from the futures price. The consumer expectation is proxied for by applying a no-change forecast to the real price of crude oil and adding the expected rate of inflation, motivated by the results for gasoline price expectations in Anderson, Kellogg, Sallee, and Curtin (2011). The vector autoregressive model (VAR) expectation is constructed from the reduced-form representation of the oil market model of Kilian and Murphy (2014) estimated on the full sample. The model includes an intercept and 24 lags of the real price of oil, the growth rate of global oil production, a proxy for the change in global crude oil inventories, and a measure of the global business cycle. The implied predictions are scaled as in Baumeister and Kilian (2014) and converted to dollar terms using the same expected rate of inflation as that underlying the consumer forecast. The inflation forecasts are constructed using the fixed-coefficient gap model proposed in Faust and Wright (2014).

consumer's perspective, somewhat smaller from the financial market perspective, and smallest when viewed through the lens of the vector autoregressive (VAR) model of the oil market. In fact, policymakers' oil price expectations often are close to consumers' expectations. This finding is interesting because it suggests that heterogeneity in oil price expectations and hence in oil price shocks across economic agents may matter for the transmission of oil price shocks, a possibility that has not been considered in existing research.

Figure 3 also illustrates another important point, which is that it does not take large positive oil price shocks to generate a sustained increase in the price of oil.

The most persistent surge in the price of oil in modern history occurred between 2003 and mid-2008, as illustrated in Figure 1, yet none of the positive oil price shocks between 2003.II and 2008.I shown in Figure 3 exceeded one standard deviation of the oil price shock series.

Why Is It So Difficult to Anticipate Oil Price Fluctuations?

Although it may seem that economists, policymakers, or financial market participants should be able to form accurate expectations about the future price of oil, if they actually understand the determinants of past oil price fluctuations, Figure 3 illustrates that this is not necessarily the case. The reason is that the price of oil will only be as predictable as its determinants, even if economic models of the global oil market are approximately correct. Unless we can foresee the future evolution of these determinants, surprise changes in the price of oil driven by unexpected shifts in oil demand or oil supply will be inevitable. This problem arises whether we formally model the key determinants of the price of oil, as in oil market vector autoregressive (VAR) models, or whether we rely on an intuitive understanding of oil markets as financial market participants are prone to.

The Role of Unexpected Demand Shifts Associated with the Global Business Cycle

Economic models of oil markets imply that the price of oil, all else equal, depends on the state of the global economy. This fact does not make forming oil price expectations any easier, however, because in practice these expectations can be only as accurate as our predictions of the evolution of the global business cycle. The problem is that changes in global real economic activity can be predicted at best at short horizons and even then only imprecisely. For example, empirical studies have documented that the predictive accuracy of vector autoregressive (VAR) models of the global oil market improves during times of persistent and hence predictable economic expansions or contractions, but is greatly reduced during normal times (Baumeister and Kilian 2012, 2015b).

The difficulty of forecasting the state of the global economy is illustrated by the oil market data since 2003. It is now widely accepted that the surge in the price of oil starting in 2003 was caused primarily by increased demand for crude oil from emerging Asia and notably China, as these countries industrialized on a large scale, yet Kilian and Hicks (2013) document that professional forecasters of real GDP systematically underestimated the extent of Chinese growth time and again for a period lasting five years. This example shows that not only econometric models, but also professional forecasters have limited ability to predict the state of the global economy. One of the difficulties in assessing the prospects for the Chinese economy even today is that we do not know to what extent Chinese growth after 2003 reflected a permanent structural transformation of the economy and to what extent it was fueled by expansionary macroeconomic policies that are not sustainable. Another challenge is the lack of reliable and timely data for the Chinese economy.

The Role of Unexpected Shifts in Global Oil Production

Another key determinant of the price of oil is global oil production. There will always be oil supply disruptions due to political events in oil-producing countries that are largely unpredictable. A case in point is the disruption of oil production caused by the Libyan uprising in 2011 and the subsequent civil war in Libya. Similarly, the task of predicting Iraqi oil production is complicated by the activities of ISIS and by sectarian violence, neither of which are easily predictable.

It is not just unexpected disruptions of oil production we need to be concerned with, however. An even more important task is to gauge the response of global oil production to surges in the price of oil driven by increased demand for crude oil. For example, the unprecedented 1973/74 oil price increase was followed by an equally unprecedented search for new oil fields, which resulted in substantial increases in non-OPEC oil production in the early 1980s. It was by no means clear in the 1970s how successful this search for more oil would be or how long it would take to succeed. Hence, it is not unreasonable to view the oil production increases of the early 1980s as unexpected oil supply increases or oil supply shocks from the point of view of oil market participants.

A similar situation arose as the result of the demand-driven oil price surge between 2003 and mid-2008. Indeed, there was no shortage of skeptics who doubted the ability of oil producers to satisfy increased demand even in the long run. For example, proponents of the peak oil hypothesis insisted that global oil production had permanently peaked by 2007 or that the peak was imminent.[5] A case in point is an IMF study by Benes et al. (2015) that, using data up to 2009, predicted a near doubling of the price of oil over the coming decade based on the view that geological constraints would win out over technological improvements in conserving oil use and in oil extraction.

There is little doubt that the peak oil hypothesis, taken literally, cannot be right because it ignores the fact that at higher oil price levels, more oil production will be forthcoming, as more expensive extraction technologies become profitable. For example, deep sea oil drilling becomes profitable only at sufficiently high oil prices. Yet, this hypothesis also contains a grain of truth in that as of 2008 no one would have known for sure whether future oil production would be sufficient to meet demand at current prices. Even granting that such a supply response did occur following the 1973/74 oil price surge with a delay of five years or more, the obvious question in 2008, as in any similar situation, was whether this time would be different. Nothing in past experience guaranteed as of 2008 that the oil supply response would be adequate going forward or that it would occur in a timely manner. For example, the rapid growth of US shale oil production after 2008, which was facilitated in important part by technological innovations in oil drilling, was a surprise to many analysts.

[5] The peak oil hypothesis originates with Hubbard (1956) and postulates a bell-shaped curve intended to describe the rate at which crude oil is extracted over time. Once the peak of this curve has been reached, the rate of oil production will decline permanently. This curve may be estimated from past oil production data. For an economic perspective on the peak oil debate see Holland (2008, 2013).

Looking back at this episode, we now know with the benefit of hindsight that the market for oil worked once again. As in the 1970s, it took about five years from the peak in the price of oil for the market to generate substantial increases in oil production on a global scale. This does not mean that increased scarcity will not become a reality in the long run. In this regard, Hamilton (2013) documented that historically higher oil production usually reflected the development of oil fields in new locations, rather than increased efficiency in oil production. The current US shale oil boom, which is driven by improved technology for horizontal drilling and fracking, is a counterexample. This boom is unlikely to last forever, however, even granting that efficiency in shale oil production gains may extend the length of the boom (Kilian 2014a).

The real question thus is whether demand for oil will diminish when the price of oil ultimately rises, as firms and consumers substitute alternative fuels for oil products in the transportation sector, for example. If they do, the peak oil hypothesis will become irrelevant. Indeed, no one today is concerned about the world running out of coal, yet the "Coal Question" raised by Jevons (1866) is eerily reminiscent of the peak oil hypothesis of 2007. Jevons stressed that British coal reserves were finite and would be exhausted by the 1960s if coal consumption were to grow at the same rate as the population. His predictions proved inaccurate because coal, which at the time was the primary fuel, was replaced by oil starting in the 1920s and by other fuels in the 1970s. The question is whether, in the very long run, renewable energies will do the same to oil products.

The Role of Unexpected Shifts in Inventory Demand

One more difficulty in forming oil price expectations is unanticipated changes in perceptions about the future scarcity of oil that affect future demand for oil inventories. Such perceptions may evolve rapidly, for example, in response to geopolitical or economic crises. Thus, expectations of the price of oil that may have been perfectly reasonable at the time, may be easily rendered obsolete by unforeseen political or economic events. In fact, oil prices may change merely in response to a shift in uncertainty, reflecting precautionary demand (for example, Adelman 1993; Pindyck 2004; Kilian 2009a; Alquist and Kilian 2010).

Obviously, predicting the exact timing of specific political crises and their impact on oil markets is next to impossible, even if political analysts have no trouble identifying geopolitical hotspots. The Arab Spring is a case in point. Likewise, the timing of economic crises is difficult to anticipate. An example is the financial crisis of 2008. Moreover, predicting political crises alone is not sufficient to ensure that oil price expectations are accurate. It is important to keep in mind that oil inventory demand depends on the shortfall of expected supply compared with expected demand rather than just one side of the market. Historical evidence suggests that, in practice, shifts in inventory demand tend to arise only when geopolitical turmoil coincides with expectations of strong demand for crude oil and tight oil supplies. Geopolitical tensions alone, in contrast, will not have an effect on the price of oil as long as oil supplies are plentiful relative to

expected demand (Mabro 1998). For example, as noted earlier, persistent attacks on oil tankers in the Persian Gulf during the 1980s—at the rate of as many as 30 attacks per month—had no apparent effect on the price of oil. Thus, most oil price predictions simply ignore the possibility of future political or economic crises, except to the extent that they are already priced in at the time the prediction is made. Because crises are rare, this strategy usually works, but occasionally it may result in spectacular predictive failures.

Conclusion

Although our understanding of historical oil price fluctuations has greatly improved, oil prices keep surprising economists, policymakers, consumers, and financial market participants. Our analysis focused on the question of how large this surprise component or shock component of oil price fluctuations is. We illustrated that the timing and magnitude of oil price shocks depends on the measure of oil price expectations. The reason why economists care about oil price shocks is that these shocks affect economic decisions. One channel of transmission is the loss of discretionary income that is associated with unexpectedly higher oil prices (and hence higher gasoline prices). Consumers who are forced to commute to and from work, for example, often have little choice, but to pay higher gasoline prices, which reduces the amount of discretionary income available for other purchases. Another channel is that oil price shocks affect expectations about the future path of the price of oil. Such expectations enter into net present value calculations of future investment projects, the cash flow of which depends on the price of oil. For example, an automobile manufacturer's decision of whether to build new production facilities for a sport utility vehicle is directly affected by the price of oil, as is a household's decision which car to buy. What matters for net present value calculations is not the magnitude of the oil price surprise in the current period, but the revision to the expected path of the future price of oil which enters the cash flow. In addition, oil price shocks also affect the cash flow of earlier investment decisions by manufacturing firms. In general, ongoing projects remain profitable as long as the price exceeds the marginal cost, which depends on the current price of oil. It is even possible for higher oil prices to cause ongoing projects to be abandoned (much like a consumer may choose to scrap a gas-guzzling car in response to higher gasoline prices). A more comprehensive review of the transmission of oil price shocks that also takes account of the joint determination of the price of oil and the state of the economy is provided in Barsky and Kilian (2004) and Kilian (2008b, 2014b).

These basic transmission mechanisms have been linked to a wide range of macroeconomic outcomes including inflation, output, employment, and wages (for example, Kilian 2008, 2014). They also have implications for the debate about climate change and for environmental policies. Of particular interest is the question of how the effects of oil price shocks vary across industries, plants, and households, how long these effects last, and how they may have changed over

time. One important insight of the recent literature has been that these questions cannot be answered without taking account of the underlying causes of the oil price shock. Our analysis suggests another important direction for future research. Macroeconomic models of the transmission of oil price shocks to date have not allowed for heterogeneous oil price expectations across economic agents. Given the differences in the implied oil price shocks associated with alternative measures of oil price expectations that we documented, such distinctions may be of first-order importance for applied work.

■ *We thank Ana María Herrera and the editors for helpful discussions.*

References

Adelman, Morris A. 1993. *The Economics of Petroleum Supply*. Cambridge, MA: MIT Press.

Alhajji, A. F., and David Huettner. 2000. "OPEC and other Commodity Cartels: A Comparison." *Energy Policy* 28(15): 1151–64.

Almoguera, Pedro A., Christopher C. Douglas, and Ana María Herrera. 2011. "Testing for the Cartel in OPEC: Noncooperative Collusion or Just Noncooperative?" *Oxford Review of Economic Policy* 27(1): 144–68.

Alquist, Ron, and Lutz Kilian. 2010. "What Do We Learn from the Price of Crude Oil Futures?" *Journal of Applied Econometrics* 25(4): 539–73.

Alquist, Ron, Lutz Kilian, and Robert J. Vigfusson. 2013. "Forecasting the Price of Oil." In *Handbook of Economic Forecasting*, vol. 2, edited by Elliott, Graham, and Allan Timmermann, 427–507. Amsterdam: North-Holland.

Anderson, Soren T., Ryan Kellogg, James M. Sallee, and Richard T. Curtin. 2011. "Forecasting Gasoline Prices Using Consumer Surveys." *American Economic Review* 101(3): 110–14.

Barsky, Robert B., and Lutz Kilian. 2002. "Do We Really Know that Oil Caused the Great Stagflation? A Monetary Alternative." *NBER Macroeconomics Annual 2001*, vol. 16, pp. 137–83.

Barsky, Robert B., and Lutz Kilian. 2004. "Oil and the Macroeconomy since the 1970s." *Journal of Economic Perspectives* 18(4): 115–34.

Baumeister, Christiane, and Gert Peersman. 2013. "The Role of Time-Varying Price Elasticities in Accounting for Volatility Changes in the Crude Oil Market." *Journal of Applied Econometrics* 28(7): 1087–1109.

Baumeister, Christiane, and Lutz Kilian. 2012. "Real-Time Forecasts of the Real Price of Oil." *Journal of Business and Economic Statistics* 30(2): 326–36.

Baumeister, Christiane, and Lutz Kilian. 2014. "What Central Bankers Need to Know about Forecasting Oil Prices." *International Economic Review* 55(3): 869–89.

Baumeister, Christiane, and Lutz Kilian. 2015a. "A General Approach to Recovering Market Expectations from Futures Prices with an Application to Crude Oil." Unpublished paper, University of Michigan. http://www-personal.umich.edu/~lkilian/bk4_081915withappendix.pdf.

Baumeister, Christiane, and Lutz Kilian. 2015b.

"Forecasting the Real Price of Oil in a Changing World: A Forecast Combination Approach." *Journal of Business and Economic Statistics* 33: 338–51.

Baumeister, Christiane, and Lutz Kilian. Forthcoming. "Understanding the Decline in the Price of Oil since June 2014." *Journal of the Association of Environmental and Resource Economists.*

Benes, Jaromir, Marcelle Chauvet, Ondra Kamenik, Michael Kumhof, Douglas Laxton, Susanna Mursula, and Jack Selody. 2015. "The Future of Oil: Geology versus Technology." *International Journal of Forecasting* 31(1): 207–21.

Bodenstein, Martin, Luca Guerrieri, and Lutz Kilian. 2012. "Monetary Policy Responses to Oil Price Fluctuations." *IMF Economic Review* 60(4): 470–504.

Cairns, Robert D., and Enrique Calfucura. 2012. "OPEC: Market Failure or Power Failure?" *Energy Policy* 50: 570–80.

Carter, Colin A., Gordon C. Rausser, and Aaron D. Smith. 2011. "Commodity Booms and Busts." *Annual Review of Resource Economics* 3: 87–118.

Colgan, Jeff D. 2014. "The Emperor Has No Clothes: The Limits of OPEC in the Global Oil Market." *International Organization* 68(3): 599–632.

Dvir, Eyal, and Kenneth S. Rogoff. 2010. "Three Epochs of Oil." Unpublished manuscript, Boston College, August 16. http://scholar.harvard.edu/files/rogoff/files/three_epochs_of_oil.pdf.

Fattouh, Bassam, Lutz Kilian, and Lavan Mahadeva. 2013. "The Role of Speculation in Oil Markets: What Have We Learned So Far?" *Energy Journal* 34(3): 7–33.

Faust, Jon, and Jonathan H. Wright. 2013. "Forecasting Inflation." In *Handbook of Economic Forecasting*, vol. 2, edited by Elliott, Graham, and Allan Timmermann, 2–56. Amsterdam: North-Holland.

Green, Edward J., and Robert H. Porter. 1984. "Noncooperative Collusion under Imperfect Price Information." *Econometrica* 52(1): 87–100.

Hamilton, James D. 1983. "Oil and the Macroeconomy since World War II." *Journal of Political Economy* 91(2): 228–48.

Hamilton, James D. 1985. "Historical Causes of Postwar Oil Shocks and Recessions." *Energy Journal* 6(1): 97–116.

Hamilton, James D. 2003. "What Is an Oil Shock?" *Journal of Econometrics* 113(2): 363–98.

Hamilton, James D. 2009. "Causes and Consequences of the Oil Shock of 2007–08." *Brookings Papers on Economic Activity*, 1, Spring, pp. 215–283.

Hamilton, James D. 2013. "Oil Prices, Exhaustible Resources and Economic Growth." In *Handbook on Energy and Climate Change*, edited by Roger Fouquet, 29–57. Cheltenham, UK: Edward Elgar.

Hamilton, James D., and Jing Cynthia Wu. 2014. "Risk Premia in Crude Oil Futures Prices." *Journal of International Money and Finance* 42: 9–37.

Holland, Stephen P. 2008. "Modeling Peak Oil." *Energy Journal* 29(2): 61–80.

Holland, Stephen P. 2013. "The Economics of Peak Oil." In *Encyclopedia of Energy, Natural Resource, and Environmental Economics*, vol. 1, edited by Jason F. Shogren, 146–50. Amsterdam: Elsevier.

Hubbert, M. King. 1956. "Nuclear Energy and the Fossil Fuels." *American Petroleum Institute Drilling and Production Practice, Proceedings of Spring Meeting*, San Antonio, 7–25.

Jevons, William S. 1886. *The Coal Question*. London: Macmillan and Co.

Kilian, Lutz. 2008a. "Exogenous Oil Supply Shocks: How Big Are They and How Much Do They Matter for the U.S. Economy?" *Review of Economics and Statistics* 90(2): 216–40.

Kilian, Lutz. 2008b. "The Economic Effects of Energy Price Shocks." *Journal of Economic Literature* 46(4): 871–909.

Kilian, Lutz. 2009a. "Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market." *American Economic Review* 99(3): 1053–69.

Kilian, Lutz. 2009b. "Comment on 'Causes and Consequences of the Oil Shock of 2007–08' by James D. Hamilton." *Brookings Papers on Economic Activity*, 1, Spring, pp. 267–78.

Kilian, Lutz. 2014a. "The Impact of the Shale Oil Revolution on U.S. Oil and Gasoline Prices." CEPR Discussion Paper 10304.

Kilian, Lutz. 2014b. "Oil Price Shocks: Causes and Consequences." *Annual Review of Resource Economics* 6: 133–54.

Kilian, Lutz, and Bruce Hicks. 2013. "Did Unexpectedly Strong Economic Growth Cause the Oil Price Shock of 2003–2008?" *Journal of Forecasting* 32(5): 385–94.

Kilian, Lutz, and Daniel P. Murphy. 2012. "Why Agnostic Sign Restrictions Are Not Enough: Understanding the Dynamics of Oil Market VAR Models." *Journal of the European Economic Association* 10(5): 1166–88.

Kilian, Lutz, and Daniel P. Murphy. 2014. "The Role of Inventories and Speculative Trading in the Global Market for Crude Oil." *Journal of Applied Econometrics* 29(3): 454–78.

Kilian, Lutz, and Thomas K. Lee. 2014. "Quantifying the Speculative Component in the Real Price of Oil: The Role of Global Oil Inventories." *Journal of International Money and Finance* 42: 71–87.

Knittel, Christopher R., and Robert S. Pindyck. Forthcoming. "The Simple Economics of Commodity Price Speculation." *American Economic Journal: Macroeconomics.*

Lippi, Francesco, and Andrea Nobili. 2012. "Oil and the Macroeconomy: A Quantitative Structural Analysis." *Journal of the European Economic Association* 10(5): 1059–83.

Mabro, Robert. 1998. "OPEC Behavior 1960–1998: A Review of the Literature." *Journal of Energy Literature* 4(1): 3–27.

Pindyck, Robert S. 2004. "Volatility and Commodity Price Dynamics." *Journal of Futures Markets* 24(11): 1029–47.

Ramey, Valerie A., and Daniel J. Vine. 2011. "Oil, Automobiles, and the U.S. Economy: How Much Have Things Really Changed?" *NBER Macroeconomics Annual* 2010, vol. 25, pp. 333–68.

Seymour, Ian. 1980. *OPEC: Instrument of Change.* London: MacMillan.

Skeet, Ian. 1988. *OPEC: Twenty-Five Years of Prices and Politics.* New York: Cambridge University Press.

Smith, James. 2005. "Inscrutable OPEC? Behavioral Tests of the Cartel Hypothesis." *Energy Journal* 26(1): 51–82.

Yergin, Daniel. 1992. *The Prize: The Epic Quest for Oil, Money, and Power.* New York: Simon and Schuster.

Journal of Economic Perspectives—Volume 30, Number 1—Winter 2016—Pages 161–184

Using Natural Resources for Development: Why Has It Proven So Difficult?

Anthony J. Venables

Using natural resources to promote economic development sounds straightforward. A country has subsoil assets such as hydrocarbons and minerals, which it seeks to transform into surface assets—human and physical capital—that can be used to support employment and generate economic growth. Such assets should be particularly valuable for capital-scarce developing countries, especially as revenues from their sale accrue largely in foreign exchange and can supplement the otherwise limited fiscal capacity of their governments.

In practice, this transformation has proved hard. Indeed, few developing economies have been successful with this approach, and economic growth has generally been lower in resource-rich developing countries than in those without resources. It was not until the 2000s (a period of rising commodity prices) that resource-rich countries grew faster, although even then per capita growth was similar in both groups of countries (IMF 2012b). The term "resource curse" was coined (Auty 1993) to capture the underperformance of resource-rich economies, drawing attention to the weak performance of Bolivia, Nigeria, and Venezuela, amongst others.

Successful use of nonrenewable natural resources involves multiple stages. Resource deposits have to be discovered and developed. If and when this is done, resource revenues are divided between investors, government, and other claimants. How are the terms of this division decided, and how are such revenues utilized by the recipients? There is likely to be intense pressure for current spending rather than investment in assets that will be productive over time. Investment in the domestic

■ *Anthony J. Venables is BP Professor of Economics Department of Economics, University of Oxford, Oxford, United Kingdom. His email address is tony.venables@economics.ox.ac.uk.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.1.161

doi=10.1257/jep.30.1.161

economy needs to be directed to high social return projects, but these may be difficult to identify and to implement. Placing the revenues in offshore funds may be appropriate for capital-rich economies, but does little to boost economic development in a capital-poor country. Ultimately, it is the private sector that will create the sustainable jobs and economic growth, so resource management has to be done in a manner that will support private sector investments. But even if revenues are effectively utilized, resource exports can appreciate the exchange rate and prove damaging to other tradable sectors of the economy—the so-called "Dutch disease" effect. An economy with substantial exports of natural resources can become overly dependent on a single volatile source of income, and this volatility can destabilize the macroeconomy.

Subsoil assets are property of the state in almost all countries except the United States. Thus, to navigate the multiple stages in the use of natural resources successfully, governments in resource-rich countries need to be well-intentioned, far-sighted, and highly capable. Yet many resource-rich economies have weak governance that can be further undermined by the political forces that are unleashed with the prospect of resource wealth.

The multistage nature of the challenge means that no single answer can be given to the question of why it has proven so difficult to harness natural resource wealth for broader economic development. While some countries have succeeded in using natural resources for development, others have failed, each in their own way. This paper discusses the challenges posed by each of these stages, the evidence on country performance, and some particular country examples. We start by outlining the scale of the issue and the main facts about resource-rich low-income countries. Following sections then turn to each of the main stages: the upstream issue of attracting investment in the resource sector and securing a flow of resource income; the economics and politics of managing revenue from natural resources; and the wider impact of substantial natural resource exports on the structure and diversification of the economy. Lessons in all of these areas, along with the future prospects for resource-rich low-income countries, can be drawn both from resource-rich countries that have succeeded in building on their resource base and from those which have not.

Facts

The IMF classifies 51 countries, home to 1.4 billion people, as "resource-rich." This classification is based on a country deriving at least 20 percent of exports or 20 percent of fiscal revenue from nonrenewable natural resources (based on 2006–2010 averages as explained in IMF 2012b). In 25 of these countries, resources make up more than three-quarters of exports, and in 20 of them resources provide more than half of government revenues. A full list of the 51 countries, along with a further 12 developing countries that are "prospectively" resource rich, is available in the online Appendix available with this paper at http://e-jep.org. The

Table 1
Resource Dependent Low- and Lower-Middle-Income Countries

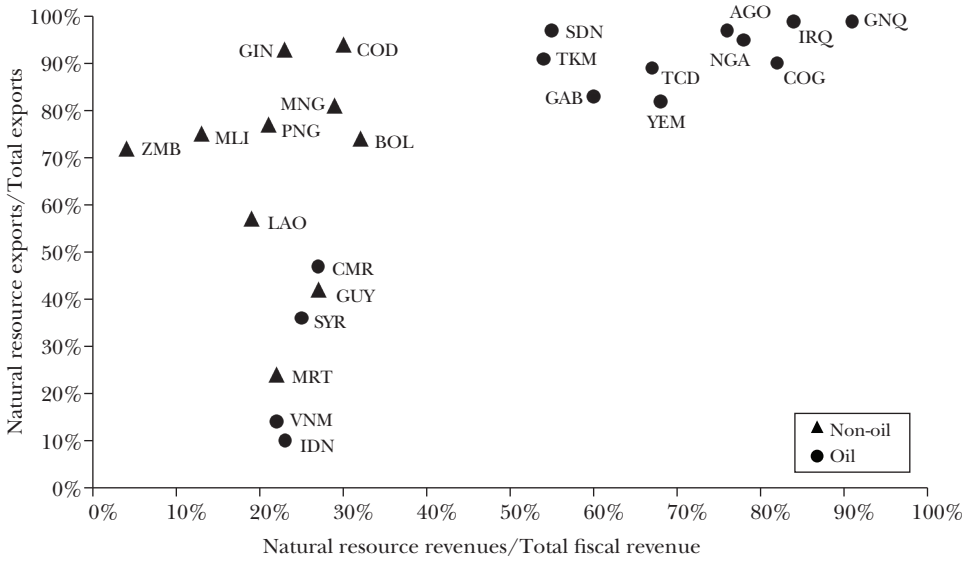
Country	*Type of natural resource*	*GNI per capita (2010 US$)*	*Natural resource exports as % of total exports (2006–2010 average)*	*Natural resource fiscal revenue as % of fiscal revenue (2006–1000) average)*
Congo, Dem. Rep.	Minerals & Oil	180	94	30
Liberia	Gold & Iron Ore	210	—	16
Niger	Uranium	360	—	—
Guinea	Mining Products	390	93	23
Mali	Gold	600	75	13
Chad	Oil	710	89	67
Mauritania	Iron Ore	1,000	24	22
Lao PDR	Copper & Gold	1,010	57	19
Zambia	Copper	1,070	72	4
Vietnam	Oil	1,160	14	22
Yemen	Oil	1,160	82	68
Nigeria	Oil	1,170	97	76
Cameroon	Oil	1,200	47	27
Papua New Guinea	Oil/Copper/Gold	1,300	77	21
Sudan	Oil	1,300	97	55
Uzbekistan	Gold & Gas	1,300	—	—
Côte d'Ivoire	Oil & Gas	1,650	—	—
Bolivia	Gas	1,810	74	32
Mongolia	Copper	1,870	81	29
Congo, Rep. of	Oil	2,240	90	82
Iraq	Oil	2,380	99	84
Indonesia	Oil	2,500	10	23
Timor Leste	Oil	2,730	99	—
Syrian Arab Rep.	Oil	2,750	36	25
Guyana	Gold & Bauxite	2,900	42	27
Turkmenistan	Oil & Gas	3,790	91	54
Angola	Oil	3,960	95	78
Gabon	Oil	7,680	83	60
Equatorial Guinea	Oil	13,720	99	91

Source: World Development Indicators, World Bank; and IMF staff estimates.

upper-middle-income resource-rich economies are a mixed group, including countries from Latin America (like Chile and Venezuela), central Asia (Azerbaijan and Kazakhstan), and Africa (Libya and Algeria). The high-income resource-rich economies are mainly Middle Eastern oil exporters, along with Norway and Trinidad and Tobago. Of the twelve "prospectively" resource-rich countries, with new discoveries that are yet to be fully developed, nine are in Africa.

Our focus is on low- and lower-middle-income resource-rich countries. There are 29 such countries, which are listed in Table 1. For this group there are four key facts. First, for many of these countries, there is extreme dependence on natural

Figure 1
Share of Exports and Fiscal Revenue from Natural Resources
(average 2006–2010)

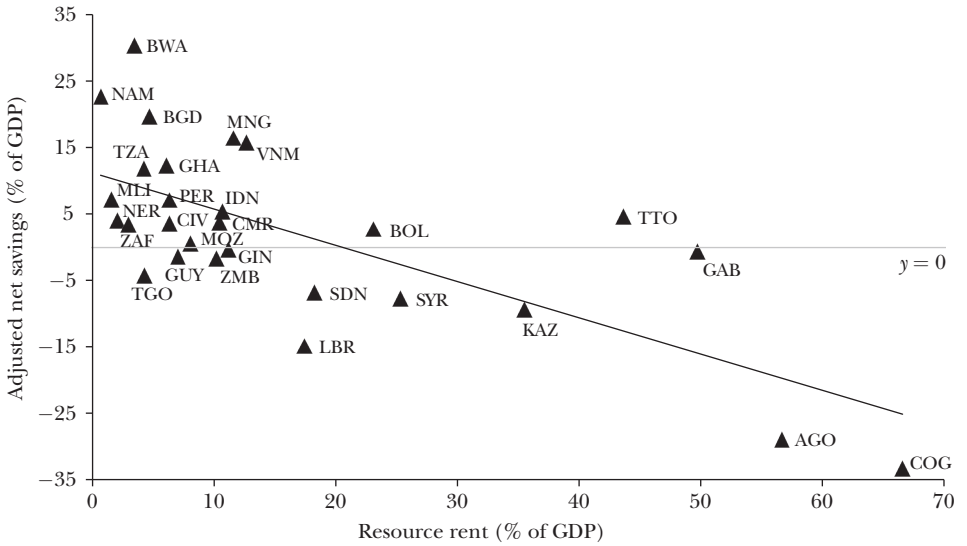


Sources: World Development Indicators, World Bank; and IMF staff estimates.
Notes: AGO = Angola; BOL = Bolivia; CMR = Cameroon; COD = The Democratic Republic of Congo; COG = Republic of the Congo; GAB = Gabon; GIN = Guinea; GNQ = Equatorial Guinea; GUY = Guyana; IDN = Indonesia; IRQ = Iraq; LAO = Laos; MNG = Mongolia; NGA = Nigeria; MLI = Mali; MRT = Mauritania; PNG = Papua New Guinea; SDN = Sudan; SYR = Syria; TCD = Chad; TKM = Turkmenistan; VNM = Vietnam; YEM = Yemen; ZMB = Zambia.

resources for fiscal revenues, export sales, or both. Figure 1 plots the fiscal and export dependency of the 24 of these countries for which reliable data are available. Ten of them receive more than half of fiscal revenue from resources, and in 17 of these countries, resources constitute more than two-thirds of their exports. Fiscal dependency is particularly acute for oil producers.

Second, saving in these low-income resource-rich economies has generally been low. This is illustrated in Figure 2, showing the relationship between resource rents and adjusted net savings, both expressed as a percentage share of GDP, for 28 middle- and low-income resource-rich countries. Resource rents are measured by the World Bank in its World Development Indicators as gross revenues from oil, natural gas, coal, minerals, and forests minus their estimated extraction costs. Adjusted net savings are national savings plus education expenditure and minus depletion of natural resources (World Bank 2011). As is apparent, this measure of adjusted national saving is strongly negative for a large number of resource-rich low-income economies, and there is a negative correlation between resource rents and the savings rate.

Figure 2
Adjusted Net Savings and Exhaustible Resource Rent
(average 2000–2009)



Sources: World Development Indicators, World Bank; and IMF staff estimates.
Notes: AGO = Angola; BGD = Bangladesh; BOL = Bolivia; BWA = Botswana; CMR = Cameroon; COG = Republic of the Congo; CIV = Côte d'Ivoire; GAB = Gabon; GHA = Ghana; GIN = Guinea; GUY = Guyana; IDN = Indonesia; KAZ = Kazakhstan; LBR = Liberia; MNG = Mongolia; NAM = Nambia; NER = Niger; MLI = Mali; MOZ = Mozambique; PER = Peru; SDN = Sudan; SYR = Syria; TGO = Togo; TTO = Trinidad and Tobago; TZA = Tanzania; VNM = Vietnam; ZAF = South Africa; ZMB = Zambia. Resource rents are measured by the World Bank in its World Development Indicators as gross revenues from oil, natural gas, coal, minerals, and forests minus their estimated extraction costs. Adjusted net savings are national savings plus education expenditure and minus depletion of natural resources.

Third, the growth performance of all the resource-rich economies as a group has been generally poor, although a few countries have done well—for example, Botswana, Malaysia, and Chile. This cross-country finding has been extensively researched following the seminal work of Sachs and Warner (1995, 1997) who found (after controlling for initial income per capita, investments in physical and human capital, trade openness, and rule of law) that natural resource dependence had a significant negative effect on the growth of GDP per capita, with a 10 percentage point increase in the ratio of resource exports to GDP depressing average growth by 0.77–1.1 percentage points per annum. Important later contributions include Mehlum, Moene, and Torvik (2006) who interact resource abundance with institutional quality and find the negative effect of resource-richness on growth to be present (and larger) only for countries with poor institutional quality, the break-even point being around the institutional quality of Botswana. More recent work has looked at some other dimensions of the connection from natural resource wealth to growth. For example, subnational evidence finds that the local impact

of extraction has positive effects (Cust and Poelhekke 2015), but the local impact of rent distribution is negative (Caselli and Michaels 2013). An extensive review of this literature, which also discusses the endogeneity issues associated with different measures of resource abundance, is found in Smith (2015).

Looking just at developing countries, there has been a recent improvement in the relative performance of resource-rich economies, with average per capita growth rates of resource-rich developing economies equalling those of nonresource rich in the 2000s, after being 1 percent per year lower in the 1990s. Of course, much of the earlier 2000s was also a time of booming oil and commodity prices and of rising resource trade with China, so this remains a very modest growth performance. As Ross (2012) wrote of growth performance of resource-rich economies: "[T]he real problem is not that growth . . . has been slow when it should have been normal, but that it has been normal when it should have been faster than normal."

Fourth, resource revenues can be highly volatile. Some variability is predictable—due to opening of new deposits of natural resources and closure of depleted ones—but much is unpredictable and largely due to the volatility of commodity prices, particularly that of oil. There is a large literature on the measurement and causes of commodity price instability (for example, Arezki, Loungani, van der Ploeg, and Venables 2014), and our concern is principally its impact on resource producers. The scale of the issue is vividly illustrated by the fact the World Bank's measure of resource rents, for the world as a whole, has fluctuated at between 1½ percent (1998) and 7 percent (2008) of world GDP over the last 20 years. Amongst resource-rich developing economies, measures of volatility (for example, the coefficient of variation of export revenues) typically exceed those of nonresource-rich countries by 50 percent for mineral-rich countries and more than 100 percent for oil-rich countries. Smoothing is made difficult by the long cycles of many commodity prices (particularly oil, elevated in the periods 1974–85 and 2003–2014 and with long periods of lower, but still variable, prices in between). Volatility of fiscal revenues is transmitted into even greater volatility of government spending as a consequence of procyclical public spending (IMF 2012a, b). A study by van der Ploeg and Poelhekke (2009) decomposes the effect of resource dependence on growth into a direct and a volatility effect, finding that the direct effect is positive but often dominated by the negative indirect effect through volatility.

Discovery, Development, and Rent Capture

Prerequisites for using natural resources to promote economic development are their discovery, investment in the mines or wells necessary for their extraction, and securing the subsequent flow of income. These upstream stages of resource management are complex, and the resource endowments of many developing countries remain underexplored and underexploited.

Initial discovery and development of a natural resource deposit requires investment by firms with considerable technical expertise. In developing countries,

these firms are generally foreign-owned. Economic principles suggest that the host country—owner of the resource—should put in place a regulatory and fiscal regime in which the investor can make a normal rate of return, and rents over and above this rate can then be captured by the resource owner, the state. A regime of this sort has a number of elements. Exploration and development licenses generally carry a fee, often determined by auctioning of the rights. Subsequent resource extraction is taxed through a combination of royalties on output, production-sharing agreements in which a certain fraction of production is taken by the government directly, and through corporate income tax, possibly at a rate specific to the extractive sector. Actual practice varies widely between countries, but one straightforward example is the sale of US oil and gas exploration and development rights on the outer continental shelf between 1983 and 2002. Sale was by first-price sealed bid and raised $16 billion from fees (bonus payments) on winning bids, and a further $14 billion from subsequent royalties on the 15 percent of tracts where exploration was successful and production took place (as measured in 1982 prices, according to Hendricks and Porter 2014). The Libyan auction of 2005 offers a more complex example. Investors bid a production share and other terms with, for example, one particular winning bid giving government 88 percent of gross revenue; government paying 88 percent of operating costs and lower shares of exploration and development costs; and, once cost recovery is complete, the company's profits on the remaining 12 percent being subject to a tax rate rising from 10 to 50 percent (Cramton 2010).

Even this quick sketch of the regulatory and tax problem suggests a number of complicating factors that can deter investors and depress the revenues that can be captured by the state. First, the process through which licenses are allocated can raise difficulties. Ideally, this process is transparent, competitive, and can secure a high fraction of the rent for the state. Auctions will often be useful, but are not appropriate in all cases: for example, where there is a single dominant bidder. Thus, Botswana negotiated rights to diamond extraction with dominant player De Beers, rather than using an auction. The use of auctions is now widespread (particularly for oil, less so for hard-rock minerals), but there are many instances of rights having been awarded in ways that are nontransparent and possibly corrupt, and thus not ending up with the best-qualified investor. A recent example involves the Simandou iron-ore project in Guinea (*The Economist* 2014).

Second, investments in discovery and extraction of nonrenewable resources are inherently risky due to geological and price uncertainty. Investors are further deterred by uncertainty surrounding the local economic, institutional, and political environment. The regulatory environment may be cumbersome and unpredictable. Weak infrastructure may increase extraction costs. Security may be a concern, and the resource itself may be subject to theft. In Nigeria, theft of crude oil (known as "bunkering") is estimated to run at 10–15 percent of total production (Katsouris and Sayne 2013; Council on Foreign Relations 2015). Theft also occurs through corruption in award of contracts, as in the Petrobras scandal that is shaking Brazil (*The Economist* 2015).

Added to this, investors may be deterred by risk of hold-up. Investments are sunk and long-lived, and governments, present and future, will have an incentive to change contractual and fiscal terms once the investment is in place. At the extreme, there is expropriation risk, but there is a broader risk of changes in rates of taxation and tax allowances. This incentive is countered by reputational risk the government faces if it expects to develop future fields and, in some cases by a variety of legal mechanisms. Bilateral investment treaties offer investors protection against breach of contract. Where such treaties do not exist, countries can offer contract-specific stabilization agreements that guarantee terms (or equivalent value), the credibility of which can be reinforced by offering international arbitration and waiving sovereign immunity. Some of these agreements have been breached (as in the Zambian example to be discussed shortly), but legal remedy has rarely been sought by investors, since this path would severely damage the investor's relationship with the host country. Nevertheless, such agreements are judged to have offered some security to investors, principally by steering countries that have experienced changed circumstance towards contract renegotiation rather than unilateral action (Daniel and Sunley 2010).

Other inefficiencies arise in the regime for taxing output. Ideally, the regime should tax rents, leaving marginal extraction decisions unaffected. However, investors can disguise profits by accounting practices such as transfer mispricing (of inputs and, for specialty minerals, also of outputs). A response is to tax observable outputs, which in practice means to use royalties and production sharing agreements, even though these methods are inefficient since they distort investment and extraction decisions (Mullins 2010). The tax regime also determines the time profile of revenues and risk-sharing between government and investors. Government impatience and risk aversion militate towards the use of royalties and production-sharing agreements rather than a pure profits tax.

What are the implications of these difficulties in the relationship between investor and host country? In some cases, government "take" (that is, the share of revenues) has been exceptionally low. An example is the Zambian copper industry which, following an unsuccessful privatization, was resold with a fiscal regime that was equivalent to an effective royalty rate of 0.6 percent, one-tenth that of comparable mining projects (Adam and Simpasa 2011). These fiscal terms turned out to be unsustainable and were revised in 2008, breaching the fiscal stability assurances that had been given, but no action was brought against the government (Daniel and Sunley 2010).

Response to low take—or more generally, to the dominant role of foreign investors—has led to "resource nationalism," including the development of national resource companies to work with, or in some cases to take over, foreign investors. In the oil sector, the formation of such national oil companies occurred largely in the 1970s, and such firms now control 90 percent of world oil reserves and over 70 percent of production. The experience of these companies has been mixed. Some of them have attained world-class efficiency levels, like Saudi-Aramco and Petronas of Malaysia. Others have failed to provide effective management, in some

cases leading to dramatic declines in output, like the Nigerian National Petroleum Corporation and Zambian Consolidated Copper Mines. McPherson (2010) details further country experiences.

The more widespread problem has not been low government take from resources that are discovered and developed, but rather a failure to undertake exploration and the follow-up investments. The deterrent effect of weak institutions is studied by Cust and Harding (2013), who look at investment in areas with similar geology on either side of an international border. They find that lower institutional quality (a one-standard-deviation reduction in the political rights index produced by Freedom House) halves the number of wells drilled. A sharp example is the Albert Graben geological basin between Uganda and the Democratic Republic of the Congo, where all exploration (and substantial discoveries) has been on the Ugandan side. The scale of the problem is indicated by Collier (2007), who estimates that the value of known subsoil assets per square kilometer in Africa is just one-quarter of those remaining in OECD countries, which seems to be most likely a consequence of lack of exploration rather than resource-barren geology in Africa.

Managing Revenues

Despite these difficulties, many countries derive a high share of fiscal revenues from the natural resource sector (as shown earlier in Figure 1). What principles should guide the use of such revenues, how well have those principles been followed in practice, and why the divergence between principles and practices?

Principles

There are three key questions about the use of rents from extraction of nonrenewable resources: 1) Should the use of these resources be focused on current consumption or on investment? 2) For the investment component, what financial, physical capital, and human capital assets should be acquired? 3) Should the rents be handled by the government directly or handed to citizens? I address these questions in turn.

Concerning the question of whether revenues from nonrenewable resources should be spent on current consumption or on investment, one ethical position is that of custodianship: the current generation should pass assets on intact to future generations. In contrast, a utilitarian would argue for spreading the benefits across present and future generations. Economists' usual characterization of this approach is the permanent income hypothesis which implies that, following a discovery of an exhaustible natural resource, consumption should increase by the expected annuity value of the discovery, with revenues in excess of this being invested to build a stock of assets sufficient to finance the consumption increment in perpetuity. However, this rule needs modification in a developing economy that is capital-scarce and accumulating capital as it converges to a higher income path. The consumption increment should then be somewhat front-loaded; less should

go to future generations (who will have higher incomes in the future anyway) and more to current poverty reduction. In effect, this change brings forward (and therefore flattens) the path of consumption growth in the economy (van der Ploeg and Venables 2011). But even with this modification, the theory suggests a high savings rate from resource revenues.

Concerning the question of what assets should be acquired with resource revenues that are saved, at the aggregate level this is a choice between domestic and foreign assets. For a capital-abundant country, the usual answer is to accumulate foreign assets in a sovereign wealth fund, such as Norway's Pension Fund. For a capital-scarce country, the priority is to build domestic assets—including human as well as physical capital. Scarcity not just of capital as a whole, but of public funds in particular, suggests that government investment in infrastructure and in public health and education systems should offer high social returns. However, scarcity of funds does not automatically imply that high-return projects are immediately available. An efficient path of investment needs to take into account domestic opportunities and the absorptive capacity of the economy.

While the priority is domestic investment, there are several reasons for supporting this with some accumulation of foreign assets. One is that the efficient path of domestic investment will, quite generally, be different from the actual path of revenue, often building up more slowly and being less volatile. This suggests the need for a "parking fund"—that is, a way of placing revenues offshore until they can be used efficiently in the domestic economy. Another reason is the need to self-insure against price uncertainty by building a "stabilization fund." Some insurance against price fluctuations can be provided by financial instruments. Much oil is sold forward—that is, a price is agreed upon in the present at which the oil will be sold in the future, typically at durations up to six months. Mexico goes further, purchasing options; for example, in 2015 Mexico spent more than $1 billion to guarantee a 2016 price of least $49 a barrel on an output of 212 billion barrels of oil. However, these financial instruments are relatively short run, and so do not provide protection against the long swings of resource prices. Depositing revenues in a stabilization fund when resource prices are high is a way of building such a protective buffer.

Concerning the issue of who makes these consumption and investment decisions, the broad distinction is between the government and the private sector. The government, while distributing some revenues through current spending, can retain ownership of assets that are acquired. These may be public investments, or assets associated with lending to the private sector, perhaps through a development bank or simply by having lower government (domestic) debt than would otherwise have been the case. The alternative is that funds are given to the private sector by tax reductions or a program of citizen dividends. An example is the US state of Alaska, where fossil fuel revenues are placed in a fund, income from which is paid directly to citizens through the Permanent Fund Dividend Program.

The case for government control is derived from the scarcity of public funds in developing countries and the need to increase public investment in human

and infrastructure capital. Resource revenues can fund such investments without imposing taxes that will be distortionary and can be hard to administer in low-income countries. Furthermore, government can smooth spending, both across generations and also in response to short-run business cycle fluctuations, mitigating the risk of resource-induced macroeconomic instability. The potential benefits of distribution to the private sector are based largely on the poor track record of governments. Direct distribution to citizens may reduce the risk of corruption and improve the quality of investments undertaken, although the link from citizen dividends to efficient investment is questionable in a country with poorly developed financial institutions. Citizen dividend schemes also create their own political risks, as they may become highly politicized and subject to electoral bidding wars by populist politicians (Gupta, Segura-Ubiergo, and Flores 2014).

Outcomes

How and why do actual outcomes differ from these principles?

On the basic question of whether a significant proportion of the rents from extraction of nonrenewable resources are being saved, Figure 2 earlier showed that savings rates (in any form, whether domestic investment or foreign funds) have generally been low for low-income resource-rich countries. Public investment as a share of GDP has been (until the 2000s) lower in resource-rich low-income countries than in other low-income countries (IMF 2012b). There is cross-country panel evidence that higher resource rents are actually associated with lower public capital stocks, particularly in countries with weak institutions (Bhattacharya and Collier 2014). When governments have sought to invest savings from resource revenues, the results have often been inefficient in both design and implementation. There are numerous white elephant projects, and resource-rich countries perform poorly on the IMF's index of public investment management efficiency (Dabla-Norris, Brumby, Kyobe, Mills, and Papageorgiou 2012).

Some countries have established sovereign wealth funds in which the state invests resource revenues offshore. Botswana's Pula Fund has been successful in managing both long-run investments and stabilization. Spending by Botswana's government has been de-linked from current resource revenues, and revenues that do not meet government spending and investment criteria are invested abroad through the fund (IMF 2012b). Other experiences have been less happy. Nigeria's Excess Crude Account has played some role in stabilizing the economy, but its effectiveness has been undermined by failure of many state governments to ratify the federal Fiscal Responsibility Act that set up the fund; by absence of sound legal foundation; and by "ad hoc disbursements" (IMF 2011). Gauthier and Zeufack (2011) study the experience of Cameroon, which was initially praised for setting up an offshore (and extra-budgetary) account to manage oil revenues, but from which about half of Cameroon's total oil revenue subsequently disappeared. The overall record on stabilization funds has been poor, with multiple episodes of boom and bust. In Collier and Venables (2011), we offer a number of examples, including a study of Chile's successful Economic and Social Stabilization Fund.

Transfers of funds from the public sector to the private have been achieved to varying extents and by different means. Some of the transfer comes from lower taxes, with the average share of tax revenue in GDP being 0.2 percentage points lower for each 1 percent of GDP earned by government from resource revenues (Bornhorst, Gupta, and Thornton 2009). Citizen dividend schemes are rare in developing countries. Mongolia established a scheme, but it was scaled back dramatically in 2012 after exaggerated election promises led to transfers that exceeded resource earnings (Yeung and Howes 2015). Transfers of resource revenues to the private sector are often achieved through highly inefficient mechanisms, with fuel subsidies being the most notorious example in which oil exporters are among the highest subsidisers. For example, the price of gasoline in Venezuela has been less than $0.10 per gallon and Iran's energy subsidies peaked at 10 percent of GDP in 2010, shortly before a subsidy reform program was launched (for a broader picture, see Coady, Gillingham, Ossowski, Piotrowski, Tareq, and Tyson 2010; Cody, Parry, Sears, and Shang 2015).

A more subtle issue arises with the interaction between public saving and private sector behavior. At least some fraction of public sector saving will be perceived to ultimately accrue to the private sector—for example, leading to expectations of lower future taxes or higher pensions—which can lead to changes in private sector behavior that may undermine government policy. In Kazakhstan, the government acted prudently, saving around one-third of oil revenue in a sovereign wealth fund. But the private sector ran up foreign debt of a similar magnitude, leading to a severe crash in 2007–2008 (Esanov and Kuralbeyeva 2011). It appears that foreign borrowing by Kazakhstan's banking sector was facilitated by the perceived collateral of sovereign assets.

Causes

These examples illustrate what has gone wrong with the management of revenues from extraction of natural resources; but why have matters so often gone so wrong? Part of the answer lies in technical difficulty: coping with massive fluctuations in export earnings or with private credit booms is challenging for any government. Part is due to weak governance, which has, in some cases, been further damaged by the presence of resource revenues. Here, I will focus on issues of fiscal discipline, patronage politics, and the situations in which resource revenues inflame conflict—up to and including civil war.

Many resource-rich countries have found it difficult to maintain *fiscal discipline* in the face of competing claims for a share of resource revenues. The literature has approached this problem in various ways, the simplest of which is a model in which groups are powerful enough to obtain public spending for their projects even though the projects yield low social returns (Velasco 1999; Tornell and Lane 1999). The groups might include spending ministries, regional governors, or city mayors, all of whom have with legitimate claims on public funds. After all, it is the job of cabinet ministers or subnational government agencies to make a case for additional funding for their own departments or areas. However, since the tax base

is shared while benefits of these projects accrue disproportionately to members of a particular group, each will still bid for more funds than is efficient, even though they recognize that their own projects have low returns and displace higher-return commonly owned public assets.

The problem is exacerbated by weak government capacity. Limited capacity to appraise and implement projects means that, the larger are revenues, the greater the proportion of bad projects that get accepted. Limited capacity to police spending means that as revenues increase, corruption increases more than proportionately; the positive relationship between resource abundance and levels of corruption is established in a number of studies (for example, Ades and Di Tella 1999; Leite and Weidman 1999). More broadly, resource revenues enable government to postpone economic reforms. Normally, if a government embarks upon an economic strategy that imposes large costs across its economy, change will eventually be forced upon the government by the decline of revenue. However, resource rents provide a cushion. Chauvet and Collier (2008) find that resource rents significantly reduce the speed of exit from dysfunctional policies, as measured by a low score on the World Bank's Country Policy and Institutional Assessment (CPIA) indicator.

How might these failures of fiscal discipline be countered? Managing expectations can help. There is usually little public or even official knowledge of the actual scale of resource revenues, and there is often a tendency to overestimate wealth and ignore trade-offs. Combine these factors with individuals' uncertainty about how or when they might see benefits, and it is unsurprising that inefficient transfer mechanisms—such as fuel subsidies—become extremely hard to reverse. The implication is that transparency is important, so that revenue flows and spending are visible to parliament and civil society.

A centralized system of financial control and authority can help with fiscal discipline, too. In principle, a central finance ministry can balance the competing demands of spending ministries, regional authorities, or other lobby groups. However, to play this role effectively the finance ministry must have control of incoming revenues, along with sufficient political will and power to resist competing demands. Botswana has had a powerful Ministry of Finance and Development Planning that has controlled and prioritized spending. It recognized that, particularly after its diamond discoveries, the main constraint was not finance, but rather implementation capacity. Foreign expertise was brought into the ministry to support implementation of rigorous project appraisal and cost–benefit analyses of public spending (Criscuolo undated; Criscuolo and Palmade 2008). In many other countries control is diffuse, often with national resource companies engaging in off-budget quasi-fiscal activities, such as running fuel subsidy or even social welfare programs. An extreme example is Venezuela where, in the mid 2000s, the national oil company PDVSA was spending 40 percent more on social programs than on its oil and gas operations (McPherson 2010).

The hand of the finance ministry can be strengthened by a "fiscal constitution" that imposes ceilings on public spending from resource revenues or public funds more generally (Poterba and von Haagen 1999; Primo 2007). Many resource-rich countries

have put fiscal rules in place, assigning shares of resource revenue to different funds, some domestic and some offshore. Experience is country-specific, but overall an IMF study concluded that there is no evidence that fiscal rules have had an effect on fiscal outcomes (Ossowski, Villafuerte, Medas, and Thomas 2008). Amongst resource-rich countries, Chile's fiscal constitution has been largely successful (Frankel 2011). As a counterexample, Ghana established funds in its Petroleum Revenue Management Act of 2011 and deposited some revenues in Heritage and Stabilization funds. But strong fiscal rules governing the small resource sector coexisted with lax budget rules elsewhere, allowing government current spending to increase dramatically, creating fiscal and external deficits that necessitated an IMF rescue program early in 2015 (IMF 2015).

Spending pressures are magnified by the prevalence of *patronage politics*, which distorts public spending to favor partisan groups. This distortion can have an intertemporal dimension, with the current government spending heavily on its favored group and passing on too little capital (or too high levels of debt) to the next government (Alesina and Tabellini 1990; Alesina and Drazen 1991). Revenues can be used by the incumbent government to increase the probability of staying in power. For example, the government can initiate spending which it can credibly commit to continue if it wins the election but which the opposition party would cancel. Public sector employment in which the government hires its supporters is a good example. Robinson and Torvik (2005), and Robinson, Torvik, and Verdier (2005, 2006) show that it is possible that a substantial fraction of resource revenues are dissipated this way and, if public employment is of lower social value than the alternative, real income can be reduced by a resource windfall.

Resource politics plays out in democracies, and also enables autocrats to remain in power. Ross (2012) shows that the democratic transitions that affected many countries in the 1980s and 1990s left most oil states untouched, a finding that is not due to simply the high incidence of autocracy in the Middle East.

Wealth from natural resources can also increase *conflict risk*. As case studies (Klare 2001) and statistical analyses (Fearon and Laitin 2003; Collier and Hoeffler 2004) show, it can provide both the motive and the means for insurgency, while also providing funds for the government (or those with access to government funds) to equip itself to retain power. Besley and Persson (2008) find that an increase in commodity prices (a measure of resource revenues exogenous to each country) significantly increases the incidence of conflict. Collier, Hoeffler, and Söderbom (2004) investigate the duration of civil wars and find that a price increase of the commodities that a country exports significantly reduces the chance that a war will be settled. Dube and Vargas (2013) add an interesting twist: using regional data for Colombia, they find that higher oil prices increased conflict while increases in coffee prices had the opposite effect, possibly by increasing the value of devoting labor time to coffee production.

While actual conflict can be devastating, the threat of conflict also matters in many situations where conflict does not actually occur. Resource rents alter the leader's probability of staying in power, and hence the economic, political, and

military strategies that are pursued (Caselli and Cunningham 2009). This is evident in the responses of countries to the threat of conflict. In Malaysia, past experience of ethnic conflict led the government to commit to inclusive growth (discussed further in the following section). In Nigeria, the experience of Biafra's attempted secession in 1967 led the country to fracture into 36 separate states. Each is militarily incapable of seceding from the 35 others but, by reducing central authority, the fracture has also diminished the effectiveness of resource governance—for example, by limiting the implementation of the national Fiscal Responsibility Act.

Natural Resources and Economic Structure

Dutch Disease

Resource revenues alter the structure of the economy, particularly in countries where they constitute a share of exports at the levels indicated in Figure 1. Other tradable activities will be displaced, partly as factors of production are drawn into resource extraction, and partly as they are employed to meet increased demand for nontradables arising from domestic spending of resource revenues (Corden and Neary 1982). This phenomenon was christened the "Dutch disease," following the experience of Holland with development and export of its natural gas resources in the 1960s and 1970s. This changing structure of the economy has a counterpart in the balance of payments, as higher resource exports lead to some combination of higher imports or lower nonresource exports together with (depending on elasticities) an appreciation of the real exchange rate.

Empirical work establishes the presence of these effects. Adverse effects on nonresource tradable sectors are documented for many countries—for example, the collapse of Nigerian agriculture (Ross 2012)—and cross-country empirical work confirms that resource exports are associated with smaller tradable goods sectors. Brahmbhatt, Canuto, and Vostroknutova (2010) find that countries in which the resource sector accounts for more than 30 percent of GDP have a nonresource tradable sector 15 percentage points lower than the norm, while Ismail (2010) finds that a 10 percent increase in a measure of oil revenues is associated with an average 3.4 percent fall in value added across manufacturing.

In itself, structural change in an economy is not necessarily a problem, but it can have a negative effect on real incomes if it interacts with market failures. In particular, if the nonresource tradable sector has increasing returns (either static, or as a result of dynamic learning-by-doing), then the effect may be to reduce the level and growth of real income (Torvik 2001; Krugman 1987; Sachs and Warner 1995). Research suggests that the level and composition of exports is particularly important for economic growth (Jones and Olken 2008; Hausmann, Pritchett, and Rodrick 2005), and there is evidence that resource exports crowd out the sort of other exports that drive growth. In Harding and Venables (forthcoming), we study this by looking at the effects of resource exports on different elements of the balance of payments, finding that each $1 of resource exports typically displaces

74 cents of nonresource exports (while drawing in 23 cents of imports and having a negligible effect on the capital account). Within nonresource exports, manufacturers are more prone to crowding out than agriculture or services. Ross (2012) makes the further point that the structure of employment in resource-rich countries has had an adverse effect on women's employment opportunities and wider emancipation.

How can these adverse effects be avoided? One route is economic management to mitigate these effects and another is proactive policy to grow other sectors of the economy. We discuss each in turn.

Mitigation

Whether a resource-driven spending boom displaces other economic activity or expands activity as a whole depends on the supply response of the economy. An economy in which labor is fully employed is likely to experience a contraction of its nonresource tradable sector as employment shifts to meet expanding demand for nontradables. However, in a developing country with a substantial quantity of un- (or under)employed labor, booming demand for nontradables can draw labor into employment. This mitigates the Dutch disease and, with this increase in employment and income, the balance of payments will adjust to higher resource exports less by a reduction in nonresource exports and more by drawing in additional imports.

This mitigation is more likely to work if two conditions are met. First, the economy has to be flexible and not encounter other supply bottlenecks. This means openness to trade, ease of entry of new firms, labor market flexibility, and ease of migration to urban centres. Potential bottlenecks—such as in urban and transport infrastructure, power supply, and labor skills—need to be identified and addressed in the early stages of a resource boom, measures referred to by Collier (2010) as "investing-in-investing." Second, because these adjustments necessarily take time, spending should not ramp up too rapidly, suggesting use of a "parking fund" for resource revenues as discussed above.

A further issue arises as some economic variables may adjust faster than others—especially the exchange rate. In a flexible exchange rate regime, expectations of a future appreciation may cause an earlier appreciation, with the exchange rate jumping up at the date of resource discovery and possibly before significant spending effects are felt. The decline of tradable sectors may then precede the expansion of nontradable sectors, creating recession at least in areas of the economy not directly experiencing resource-related activity. An example is Zambia in the period 2005–2006, which experienced capital inflow due to a high nominal return on government debt and a high copper price, leading to abrupt appreciation and damage to nonresource exports (Adam and Simposa 2011). This was also part of the UK's experience with North Sea oil (Eastwood and Venables 1982). At a cross-country level, the empirical work of Arezki, Ramey, and Sheng (2015), studying the effect of giant oil discoveries, finds that the discoveries alone have an initial negative effect on employment, investment, and GDP. The appropriate

response to these expectations-driven changes is monetary and exchange rate policy that moderates upwards pressure on the exchange rate.

In summary, mitigating adverse structural change requires fiscal policies that smooth spending (and thus involve parking revenues offshore), microeconomic policies to increase the flexibility of the economy and anticipate bottlenecks, and monetary or exchange rate policies that control appreciation of the currency.

Diversification

The call for policies to grow nonresource sectors and thereby diversify the economy is widely heard, yet few resource-rich countries have been successful in doing so. What can be and has been done? Resource revenues are a source of public funds and, as is widely recommended, these can be used to fund public investments complementary to private investment, such as investment in human capital, in public infrastructure, and possibly also in utilities. As discussed above, many resource-rich economies have missed this opportunity.

Other policies can target specific sectors or firms. A frequent policy has been to promote sectors with backwards and forward linkages with the resource sector. Backward linkages arise from the resource sector's use of local inputs, and studies show that the local effects of such spending are significant, although quantitatively small (Aragón and Rud 2013; Cust and Poelhekke 2015). A number of countries have a domestic content requirement policy to strengthen these backward linkages, but such rules have generally not led to transformative growth of new activities (see *The Economist* 2015 for a discussion of Brazil's experience with Petrobras). Rigid rules are gamed, and in any case do not come free; part of any cost increase they cause is borne by the host country through reduced tax and revenue receipts. There are a few exceptions in which internationally competitive sectors have grown in this way, but the examples of the Norwegian marine engineering sector or of internationally competitive national resource companies (like Saudi Aramco or Petronas) are hard to replicate in lower-income countries. Promising new initiatives offer a more flexible approach in which natural resource firms work closely with selected local firms in order to raise their capability to qualified supplier status (for example, in meeting engineering specifications), thereby raising their potential to compete on world markets (Sutton 2014).

Forward linkages involve further processing of the natural resource either for local use or prior to export. The viability of this approach depends on the wider capabilities and comparative advantage of the local economy. Resource-rich economies have not had much success in trying to move into highly capital-intensive sectors such as petrochemicals or steel plants. However, domestic use or processing of the resource makes more sense if shipping costs are high, so there is a wedge between the world price and the domestic price. Historical transport costs meant that 19th-century economic development was often close to natural sources of coal and iron ore. In the modern economy, shipping costs are relatively low for oil and most bulk minerals, but much higher for natural gas. While the capital costs of large-scale natural gas developments (such as the offshore developments planned

in East Africa) can be met only by the prospect of export sales, the price wedge means that some fraction of output should be used domestically, which raises the important prospect of relatively cheap electricity supply for the producing region.

Governments have also pursued diversification strategies by using revenues to support investment sectors not directly linked to resources, either through development banks or direct government industrial policy. As with industrial policy in other contexts, there are numerous failures and a few successes. Malaysia offers an example of success, as does Chile. Following ethnic riots in 1969, the Malaysian government committed to using economic development to narrow racial economic inequalities (Yusof 2011). A strong central government implemented a series of development plans, a centerpiece of which was to use resource revenues (in particular oil revenues, which grew rapidly from the mid 1970s) to diversify the economy. Within agriculture, investment programs raised productivity and implemented a transition from rubber to palm oil production. In manufacturing, the economy was open to trade and foreign direct investment, and an industrial policy was pursued (including infrastructure development, particularly in special economic zones) that succeeded in developing a range of labor-intensive activities including the electronics sector. Macroeconomic stability was maintained by fiscal prudence and some element of luck, as when rapidly increasing oil volumes offset the price fall of the 1980s. Elements of Malaysia's success are due to its location in a booming region and its commodity mix (rubber and tin as well as oil). But most importantly, the government recognized that inclusive economic growth was necessary for future stability, and government capacity was sufficient to implement this policy effectively.

Concluding Comments and Future Prospects

It is straightforward to catalog the failures of resource-rich countries. Some have failed to attract investors and thus failed to receive much income from their deposits of nonrenewable resources. Many have failed to use resource revenues to finance investment at levels sufficient to support continuing nonresource growth and, with the additional impact of resource revenues on volatility and Dutch disease, other potentially dynamic sectors of the economy have failed to develop. While there is heterogeneity in country experience, underlying these symptoms are two common causes. One is the technical difficulty of handling resource revenues that are risky, volatile, and time-limited. The other is that governance has been unable to resist short-run spending pressures and commit to long-run investment and growth strategies.

What recent changes have affected the performance of resource-rich economies, and what are the future prospects?

Recent decades have seen significant improvements in aspects of governance in resource-rich countries. The quality of economic management as a whole has improved, in Africa in particular, as witnessed by improved scores on the World Bank's Country Policy and Institutional Assessment indicator, and by much

improved economic performance, with resource-rich countries growing at over 5.5 percent annually in the period 2000–2014, more than twice the rate of the 1980s and 1990s (based on data from the *World Development Indicators*). The resource sector has seen several major initiatives to improve governance. The Extractive Industries Transparency Initiative, launched in 2003, is now implemented by 48 countries, with 31 fully compliant and signed up to audit their resource revenues in a transparent manner (for background, see https://eiti.org/). Codes of best practice have been drawn up by international experts and adopted by governments and regional bodies (an example is the Natural Resource Governance Institute, http://www.resourcegovernance.org/). There is a growing realization that if resource-based spending is to be controlled successfully, there has to be not just formal processes of transparency or fiscal rules, but also citizen awareness and understanding of the possibilities and problems created by resource discoveries. Countries that have created a strong narrative of what can (and cannot) be done with resource revenues, such as Malaysia and Botswana, have found such citizen expectations to be self-fulfilling, as citizens come to see the benefits of improved economic performance, and demands for spending outside the narrative are harder to justify and easier to resist.

Improved governance, in combination with the boom in commodity prices in the first decade of the 2000s, has promoted exploration and led to new resource discoveries, notably in Africa. New players have entered resource extraction and trade, in particular China. Accompanying these changes has been the increased use of "resource for infrastructure" deals, some of which are barter deals, and others part of wider trade and investment agreements (Halland, Beardsworth, Land, and Schmidt 2014). Bräutigam and Gallagher (2014) estimate that, between 2000 and 2011, China committed $80 billion of resource-backed loans to Latin America and $53 billion to Africa—of which $13 billion is to Angola alone. The loans to Angola principally finance infrastructure, but also include school and hospital projects. Much of the construction work is done using Chinese workers and inputs and repayments are made in oil, specified in quantity, not value terms (Cassel, de Candia, and Liberatore 2010).

Such deals have potential benefits. They are a commitment to transform subsoil assets into surface assets, rather than into current consumption, and to do so in a manner that is relatively rapid. However, the devil is in the details. The terms and conditions of these contracts are generally not transparent and some appear, on close investigation, to have offered poor terms to the host economy. The quality, design, and appropriateness of projects are sometimes questionable. A 2008 agreement between the Democratic Republic of the Congo, China Exim Bank, and two Chinese construction companies worth up to $6 billion and based on giving copper and cobalt in return for infrastructure, has been criticized for lack of transparency and scrutiny, questionable project selection, and no process for assessing value for money (Global Witness 2011). To deliver their potential benefits, resource for infrastructure deals need to develop scrutiny procedures that ensure value is being derived.

Finally, future prospects for resource-rich economies are dominated by the commodity price fall of 2014–2015, viewed by some as the end of a "super-cycle" of commodity prices (for example, Goldberg 2015; Bershidsky 2015). The combination of fundamental supply-side changes in energy markets (like fracking in oil markets) and the growing efforts at conserving the use of fossil fuels in response to concerns over climate change make it likely that, at least for hydrocarbons, prices will stay low. For resource-rich countries that have been accustomed to high commodity prices in the last 10–15 years, these changes are large negative shocks. Many will have to adjust to fill two gaps, one in the public finances and the other in the balance of payments. It is to be hoped that these adjustments—increasing fiscal discipline and enabling a stronger nonresource export sector to drive growth—may improve the chances of benefiting from continuing, if reduced, revenues from extraction of nonrenewable resources.

■ *Thanks to Jim Cust, Philip Daniel, Gordon Hanson, Enrico Moretti, Timothy Taylor, and Gerhard Toews for helpful comments, and to Paul Collier and Rick van der Ploeg for many valuable discussions of the issues.*

References

Adam, Christopher, and Anthony M. Simpasa. 2011. "Copper Mining in Zambia: From Collapse to Recovery." In *Plundered Nations? Successes and Failures in Natural Resource Extraction*, edited by Paul Collier and Anthony J. Venables. London: Palgrave Macmillan.

Ades, Alberto, and Rafael Di Tella. 1999. "Rents, Competition, and Corruption." *American Economic Review* 89(4): 982–93.

Alesina, Alberto, and Allan Drazen. 1991. "Why Are Stabilizations Delayed?" *American Economic Review* 81(5): 1170–88.

Alesina, Alberto, and Guido Tabellini. 1990. "A Positive Theory of Fiscal Deficits and Government Debt." *Review of Economic Studies* 57(3): 403–414.

Aragón, Fernando M., and Juan Pablo Rud. 2013. "Natural Resources and Local Communities: Evidence from a Peruvian Gold Mine." *American Economic Journal: Economic Policy* 5(2): 1–25.

Arezki, Rabah, Prakash Loungani, Frederick van der Ploeg, and Anthony J. Venables. 2014. "Understanding International Commodity Price Fluctuations." *Journal of International Money and Finance* 42: 1–8.

Arezki, Rabah, Valerie A. Ramey, and Liugang Sheng. 2015. "News Shocks in Open Economies: Evidence from Giant Oil Discoveries." NBER Working Paper 20857.

Auty, Richard M. 1993. "*Sustaining Development in Mineral Economies: The Resource Curse Thesis.*" London and New York: Routledge.

Bershidsky, Leonid. 2015. "Maybe the Commodities Supercycle Is Actually Real." *Bloomberg View*, August 7. http://www.bloombergview.com/articles/2015-08-07/maybe-the-commodities-supercycle-is-actually-real.

Besley, Timothy J., and Torsten Persson. 2008. "The Incidence of Civil War: Theory and Evidence." NBER Working Paper 14585.

Bhattacharyya, Sambit, and Paul Collier. 2014. "Public Capital in Resource Rich Economies: Is There a Curse?" *Oxford Economic Papers* 66(1): 1–24.

Bornhorst, Fabian, Sanjeev Gupta, and John Thornton. 2009. "Natural Resource Endowment and the Domestic Revenue Effort." *European Journal of Political Economy* 25(4): 439–46.

Brahmbhatt, Milan, Otaviano Canuto, and Ekaterina Vostroknutova. 2010. "Dealing with Dutch Disease." *Economic Premise,* issue 16, World Bank.

Bräutigam, Deborah, and Kevin P. Gallagher. 2014. "Bartering Globalization: China's Commodity-Backed Finance in Africa and Latin America." *Global Policy* 5(3): 346–52.

Caselli Francesco, and Tom Cunningham. 2009. "Leader Behaviour and the Natural Resource Curse." *Oxford Economic Papers* 61(4): 628–50.

Caselli Francesco, and Guy Michaels. 2013. "Do Oil Windfalls Improve Living Standards: Evidence from Brazil." *American Economic Journal: Applied Economics* 5(1): 208–38.

Cassel, Cosima, Giuseppe de Candia, and Antonella Liberatore. 2010. "Building African Infrastructure with Chinese Money." http://www.barcelonagse.eu/tmp/pdf/ITFD10Africa.pdf.

Chauvet, Lisa, and Paul Collier. 2008. "What Are the Preconditions for Turnaround in Failing States." *Journal of Conflict Management and Peace Science* 25(4): 332–48.

Coady, David, Robert Gillingham, Rolando Ossowski, John Piotrowski, Shamsuddin Tareq, and Justin Tyson. 2010. "Petroleum Product Subsidies: Costly, Inequitable and Rising." Staff Position Note 10/05, International Monetary Fund, Washington, DC.

Coady, David, Ian Parry, Louis Sears, and Baoping Shang. 2015. "How Large Are Global Energy Subsidies." IMF Working Paper 15/105.

Collier, Paul. 2007. *The Bottom Billion: Why the Poorest Countries Are Failing and What Can Be Done About It.* Oxford University Press.

Collier, Paul. 2010. *The Plundered Planet: Why We Must—And How We Can—Manage Nature for Global Prosperity.* Oxford University Press.

Collier, Paul, and Anke Hoeffler. 2004. "Greed and Grievance in Civil War." *Oxford Economic Papers* 56(4): 563–95.

Collier, Paul, Anke Hoeffler, and Måns Söderbom. 2004. "On the Duration of Civil War." *Journal of Peace Research* 41(3): 253–73.

Collier, Paul, and Anthony J. Venables, eds. 2011. *Plundered Nations? Successes and Failures in Natural Resource Extraction*. London: Palgrave Macmillan.

Corden, W. Max, and J. Peter Neary. 1982. "Booming Sector and De-industrialisation in a Small Open Economy." *Economic Journal* 92(368): 825–48.

Council on Foreign Relations. 2015. "A Primer on Nigeria's Oil Bunkering." http://blogs.cfr.org/campbell/2015/08/04/a-primer-on-nigerias-oil-bunkering/.

Cramton, Peter. 2010 "How Best to Auction Natural Resources." In *The Taxation of Petroleum and Minerals: Principles, Problems and Practice*, edited by Philip Daniel, Michael Keen, and Charles McPherson. London and New York: Routledge.

Criscuolo, Alberto. No date. "Briefing Note: Botswana." World Bank.

Criscuolo, Alberto, and Vincent Palmade. 2008. "Reform Teams." *Public Policy for the Private Sector.* Note no. 318, World Bank.

Cust, James, and Torfinn Harding. 2013. "Institutions and the Location of Oil Exploration." Oxcarre Research Paper 127, Oxford.

Cust, James, and Steven Poelhekke. 2015. "The Local Economic Impacts of Resource Extraction." *Annual Review of Resource Economics* 7: 251–68.

Dabla-Norris, Era, Jim Brumby, Annette Kyobe, Zac Mills, and Chris Papageorgiou. 2012. "Investing in Public Investment: An Index of Public Investment Efficiency." *Journal of Economic Growth* 17(3): 235–66.

Daniel, Philip, Michael Keen, and Charles McPherson, eds. 2010. *The Taxation of Petroleum and Minerals: Principles, Problems and Practice.* New York: Routledge.

Daniel, Philip, and Emil M. Sunley. 2010. "Contractual Assurances of Fiscal Stability." In *The Taxation of Petroleum and Minerals: Principles, Problems and Practise*, edited by Philip Daniel, Michael Keen, and Charles McPherson, 405–24. New York: Routledge.

Dube, Oeindrila, and Juan F. Vargas. 2013. "Commodity Price Shocks and Civil Conflict: Evidence from Colombia." *Review of Economic Studies* 80(4): 1384–1421.

Eastwood, Robert K., and Anthony J. Venables. 1982. "The Macroeconomic Implications of

a Resource Discovery in an Open Economy." *Economic Journal* 92(366): 285–99.

Economist, The. 2014 "Crying Foul in Guinea." December 6.

Economist, The. 2015. "Whose Oil in Brazil." February 14.

Esanov, Akram, and Karlygash Kuralbeyeva. 2011. "Kazakhstan: Public Saving and Private Spending." In *Plundered Nations? Successes and Failures in Natural Resource Extraction,* edited by Paul Collier and Anthony J. Venables. London: Palgrave Macmillan.

Fearon, James D., and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97(1): 75–90.

Frankel, Jeffrey A. 2011. "How Can Commodity Exporters Make Fiscal and Monetary Policy Less Procyclical?" Chap. 10 in *Beyond the Curse: Politics to Harness the Power of Natural Resources,* edited by Rabah Arezki, Thorvaldur Gylfason, and Amadou Sy. Washington, DC: IMF.

Gauthier, Bernard, and Albert Zeufack. 2011. "Governance and Oil Revenues in Cameroon." In *Plundered Nations? Successes and Failures in Natural Resource Extraction*, edited by Paul Collier and Anthony J. Venables. London: Palgrave Macmillan.

Global Witness. 2011. "China and Congo: Friends in Need." https://www.globalwitness.org/sites/default/files/library/friends_in_need_en_lr.pdf.

Goldberg, Shelley. 2015. "The End of the Commodity Super Cycle." *Wall Street Daily*, September 1. http://www.wallstreetdaily.com/2015/09/01/commodity-prices-super-cycle.

Gupta, Sanjeev, Alex Segura-Ubiergo, and Enrique Flores. 2014. "Direct Distribution of Resource Revenues: Worth Considering?" IMF Staff Discussion Note SDN/14/05.

Halland, Håvard, John Beardsworth, Bryan Land, and James Schmidt. 2014. *Resource Financed Infrastructure: A Discussion on a New Form of Infrastructure Financing.* Washington DC: World Bank.

Harding, Torfinn, and Anthony J. Venables. Forthcoming. "The Implications of Natural Resource Exports for Non-Resource Trade." *IMF Economic Review.*

Hausmann, Ricardo, Land Pritchett, and Dani Rodrik. 2005. "Growth Accelerations." *Journal of Economic Growth* 10(4): 303–29.

Hendricks, Kenneth, and Robert H. Porter. 2014. "Auctioning Resource Rights." *Annual Review of Resource Economics* 6: 175–90.

International Monetary Fund (IMF). 2011. "Nigeria: 2010 Article IV Consultation." http://www.imf.org/external/pubs/ft/scr/2011/cr1157.pdf.

International Monetary Fund (IMF). 2012a. "Macroeconomic Policy Frameworks for Resource-Rich Developing Countries." Policy paper for the Executive Board. Washington, DC.

International Monetary Fund (IMF). 2012b. "Macroeconomic Policy Frameworks for Resource-Rich Developing Countries—Background Paper 1—Supplement 1." Washington, DC.

International Monetary Fund (IMF). 2015. "IMF Approves US$918 Million ECF Arrangement to Help Ghana Boost Growth, Jobs and Stability." Press Release No. 15/159. April 3. https://www.imf.org/external/np/sec/pr/2015/pr15159.htm.

Ismail, Kareem. 2010. "The Structural Manifestation of the 'Dutch Disease': The Case of Oil-Exporting Countries." Working Paper 10/103, International Monetary Fund, Washington, DC.

Jones, Benjamin F., and Benjamin A. Olken. 2008. "The Anatomy of Start-Stop Growth." *Review of Economics and Statistics* 90(3): 582–87.

Katsouris, Christina, and Aaran Sayne. 2013. "Nigeria's Criminal Crude: International Options to Combat the Export of Stolen Oil." Programme Report, Chatham House, London.

Klare, Michael. 2001. *Resource Wars: The New Landscape of Global Conflict.* New York: Metropolitan Books.

Krugman, Paul. 1987. "The Narrow Moving Band, the Dutch Disease, and the Competitive Consequences of Mrs. Thatcher: Notes on Trade in the Presence of Dynamic Scale Economies." *Journal of Development Economics* 27(1–2): 41–55.

Leite, Carlos, and Jens Weidmann. 1999. "Does Mother Nature Corrupt? Natural Resources, Corruption, and Economic Growth." International Monetary Fund Working Paper 99/85.

McPherson, Charles. 2010. "State Participation in the Natural Resource Sectors: Evolution, Issues and Outlook." In *The Taxation of Petroleum and Minerals: Principles, Problems and Practise,* edited by Philip Daniel, Michael Keen, and Charles McPherson. London and New York: Routledge and IMF.

Mehlum, Halvor, Karl Moene, and Ragnar Torvik. 2006. "Institutions and the Resource Curse." *Economic Journal* 116(508): 1–20.

Mullins, Peter. 2010. "International Tax Issues for the Resources Sector." Chapter 13 in *The Taxation of Petroleum and Minerals: Principles, Problems and Practise*, edited by Philip Daniel, Michael Keen, and Charles McPherson. London and New York: Routledge and IMF.

Ossowski, Rolando, Mauricio Villafuerte, Paulo A. Medas, and Theo Thomas. 2008. "Managing the Oil Revenue Boom: The Role of Fiscal Institutions." IMF Occasional Paper 260, International Monetary Fund.

Poterba, James, and Jürgen von Haagen, eds.

1999. *Fiscal Rules and Fiscal Performance*. University of Chicago Press.

Primo, David M. 2007. *Rules and Restraint: Government Spending and the Design of Institutions*. University of Chicago Press.

Robinson, James A., and Ragnar Torvik. 2005. "White Elephants." *Journal of Public Economics* 89(2–3): 197–210.

Robinson, James A., Ragnar Torvik, and Thierry Verdier. 2006. "Political Foundations of the Resource Curse." *Journal of Development Economics* 79(2): 447–68.

Ross, Michael L. 2012. *The Oil Curse: How Petroleum Wealth Shapes the Development of Nations*. Princeton: Princeton University Press.

Sachs, Jeffrey D., and Andrew M. Warner. 1995. "Natural Resource Abundance and Economic Growth." NBER Working Paper 5398.

Sachs, Jeffrey D., and Andrew M. Warner. 1997. "Sources of Slow Growth in African Economies." *Journal of African Economies* 6(3): 335–76.

Smith, Brock. 2015. "The Resource Curse Exorcised: Evidence from a Panel of Countries." *Journal of Development Economics* 116: 57–73.

Sutton, John. 2014. "Gains from the Natural Gas: Local Content and Tanzania's Industrial Development." The Seventh Gilman Rutihinda Memorial Lecture, Delivered at the Bank of Tanzania, June 10, 2014. International Growth Centre, London School of Economics, http://www.theigc.org/wp-content/uploads/2014/08/Sutton-2014-Gilman-Rutihinda-Memorial-Lecture-Speech.pdf.

Tornell, Aaron, and Philip R. Lane. 1999. "The Voracity Effect." *American Economic Review* 89(1): 22–46.

Torvik, Ragnar. 2001. "Learning by Doing and the Dutch Disease." *European Economic Review* 45(2): 285–306.

van der Ploeg, Frederick, and Steven Poelhekke. 2009. "Volatility and the Natural Resource Curse." *Oxford Economic Papers* 61(4): 727–60.

van der Ploeg, Frederick, and Anthony J. Venables. 2011. "Harnessing Windfall Revenues: Optimal Policies for Resource-Rich Developing Economies." *Economic Journal* 121(551): 1–31.

Velasco, Andrés. 1999. "A Model of Endogenous Fiscal Deficit and Delayed Fiscal Reforms." Chap. 2 in *Fiscal Rules and Fiscal Performance*, edited by James M. Poterba and Jürgen Von Hagen. Chicago University Press.

World Bank. 2011. *The Changing Wealth of Nations: Measuring Sustainable Development in The New Millennium*. World Bank: Washington DC.

World Development Indicators. No date. http://data.worldbank.org/data-catalog/world-development-indicators.

Yeung, Ying, and Stephen Howes. 2015. "Resources to Cash: A Cautionary Tale from Mongolia." Devpolicyblog, Development Policy Centre, October 22. http://devpolicy.org/resources-to-cash-a-cautionary-tale-from-mongolia-20151022/.

Yusof, Zainal Aznam. 2011. "The Developmental State: Malaysia." In *Nations? Successes and Failures in Natural Resource Extraction*, edited by Paul Collier and Anthony J. Venables. London: Palgrave Macmillan.

Journal of Economic Perspectives—Volume 30, Number 1—Winter 2016—Pages 185–206

Power Laws in Economics: An Introduction

Xavier Gabaix

Paul Samuelson (1969) was once asked by a physicist for a law in economics that was both nontrivial and true. This is a difficult challenge, as many (roughly) true results are in the end rather trivial (for example, demand curves slope down), while many nontrivial results in economics in fact require too much sophistication and rationality on the part of the agents to actually hold true in practice.[1] Samuelson answered, "the law of comparative advantage." The story does not say whether the physicist was satisfied. The law of comparative advantage is a qualitative law, and not a quantitative one as is the rule in physics. Indeed, many of the insights of economics seem to be qualitative, with many fewer reliable quantitative laws.

This article will make the case that a modern answer to the question posed to Samuelson would be that a series of power laws count as actually nontrivial and true laws in economics—and that they are not only established empirically, but also understood theoretically.[2] I will start by providing several illustrations of empirical power laws having to do with patterns involving cities, firms, and the stock

[1] In Gabaix (2014), I explore this issue by proposing a behavioral version of basic microeconomics—specifically consumer and equilibrium theory. Many nontrivial predictions of microeconomics (for example, whether demand systems have the property of Slutsky symmetry) turn out to fail when agents are boundedly rational, though more trivial predictions (for example, demand curve slope down) typically hold.

[2] At the same time, in some highly incentivized and abstract environments, other "laws" (roughly) hold, like the Black–Scholes formula, or some successful mechanisms from implementation theory.

■ *Xavier Gabaix is Martin J. Gruber Professor of Finance, Stern School of Business, New York University, New York City, New York. His email address is xgabaix@stern.nyu.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.1.185

doi=10.1257/jep.30.1.185

market. I summarize some of the theoretical explanations that have been proposed. I suggest that power laws help us explain many economic phenomena, including aggregate economic fluctuations. I hope to clarify why power laws are so special, and to demonstrate their utility. In conclusion, I list some power-law-related economic enigmas that demand further exploration.

A formal definition may be useful. A power law, also called a scaling law, is a relation of the type $Y = aX^{\beta}$, where Y and X are variables of interest, β is called the power law exponent, and a is typically an unremarkable constant. For instance, if X is multiplied by a factor of 10, then Y is multiplied by 10^{β}—one says that Y "scales" as X to the power β.

Some Empirical Power Laws

City Sizes

Let us look at the data on US cities with populations of 250,000 or greater, plotted in Figure 1. We rank cities by size of population: #1 is New York, #2 Los Angeles, and so on, using data for Metropolitan Statistical Areas provided in the *Statistical Abstract of the United States* (2012). We regress log rank on log size and find the following:

$$\ln(Rank) = 7.88 - 1.03 \ln(Size).$$

The relationship in Figure 1 is close to a straight line ($R^2 = 0.98$), and the slope is very close to 1 (the standard deviation of the estimated slope is 0.01).[3] This means that the rank of a city is essentially proportional to the inverse of its size (indeed, exponentiating, we obtain $Rank = a\, Size^{-1.03}$) with $a = e^{7.88}$). A slope of approximately 1 has been found repeatedly using data spanning many cities and countries (at least after the Middle Ages, when progress in agriculture and transport could make large densities viable, see Dittmar 2011). There is no obvious reason to expect a power law relationship here, and even less for the slope to be 1.

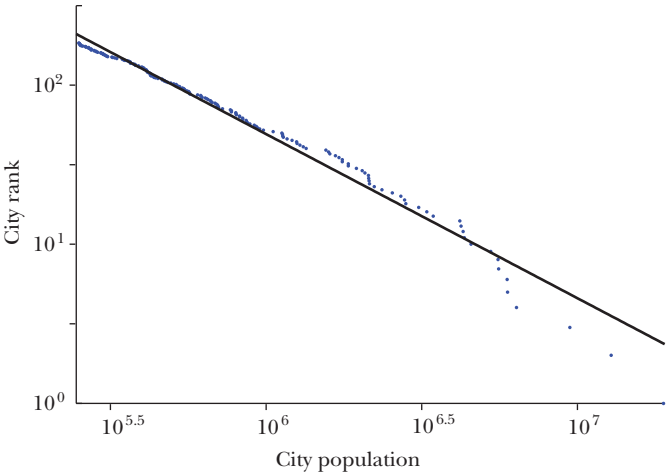
To think about this type of regularity, it is useful to be a bit more abstract and see the cities as coming from an underlying distribution: the probability that the population size of a randomly drawn city is greater than x is proportional to $1/x^{\zeta}$ with $\zeta \approx 1$. More generally,

$$\mathrm{P}(Size > x) = a/x^{\zeta},$$

at least for x above a cutoff (here, the 250,000-inhabitants cutoff used by the *Statistical Abstract of the United States*). An empirical regularity of this type is a power law. In

[3] Actually, the standard error returned by an ordinary least squares calculation is incorrect. The correct standard error is $|\text{slope}| \times \sqrt{2/N} = 0.11$, where $N = 184$ is the number of cities in the sample (Gabaix and Ibragimov 2011).

Figure 1
A Plot of City Rank versus Size for all US Cities with Population over 250,000 in 2010



Source: Author, using data from the S*tatistical Abstract of the United States* (2012).
Notes: The dots plot the empirical data. The line is a power law fit ($R^2 = 0.98$), regressing ln *Rank* on ln *Size*. The slope is −1.03, close to the ideal Zipf's law, which would have a slope of −1.

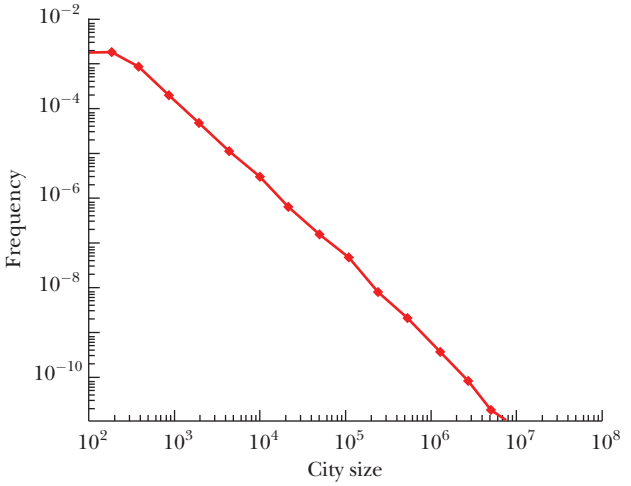
a given finitely sized sample, it generates an approximate relation of type shown in Figure 1 and in the accompanying regression equation.

The interesting part is the coefficient ζ, which is called the power law exponent of the distribution. This exponent is also sometimes called the "Pareto exponent," because Vilfredo Pareto discovered power laws in the distribution of income (as discussed in Persky 1992). A "Zipf's law" is a power law with an exponent of 1. George Kingsley Zipf was a Harvard linguist who amassed significant evidence for power laws and popularized them (Zipf 1949).

A lower ζ means a higher degree of inequality in the distribution: it means a greater probability of finding very large cities or (in another context) very high incomes.[4] In addition, the exponent is independent of the units (inhabitants or thousands of inhabitants, say). This makes it at least conceivable, a priori, that we might find a constant value in various datasets. What if we look at cities with size less than 250,000? Does Zipf's law still hold? When measuring the size of cities, it is better to look at agglomerations rather than the fairly arbitrary legal entities, but this is tricky. Rozenfeld et al. (2011) address the problem using a new algorithm that constructs the population of small cities from fine-grained geographical data. Figure 2 shows the resulting distribution of city sizes for the United Kingdom,

[4] Indeed, the expected value of S^α is mathematically infinite if α is greater than the power law exponent ζ, and finite if α is less than the power law exponent ζ. For example, if $\zeta = 1.03$, the expected size is finite, but the variance is formally infinite.

Figure 2
Density Function of City Sizes (Agglomerations) for the United Kingdom



Source: Rozenfeld et al. (2011).
Notes: We see a pretty good power law fit starting at about 500 inhabitants. The Pareto exponent is actually statistically non-different from 1 for size $S > 12{,}000$ inhabitants.

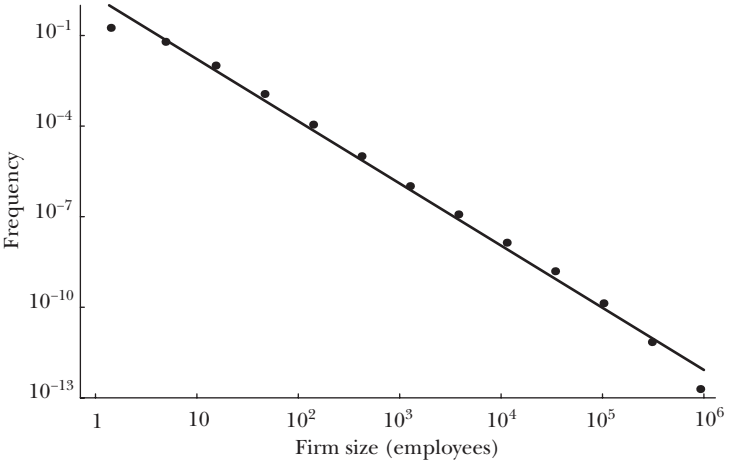
where the data is particularly good. Here we see the appearance of a straight line for cities of about size 500 and above. Zipf's law holds pretty well in this case, too.

Why might social scientists care about this relationship? As Krugman (1996) wrote 20 years ago, referring to Zipf's law, which remained unexplained by his work of economic geography: "The failure of existing models to explain a striking empirical regularity (one of the most overwhelming empirical regularities in economics!) indicates that despite considerable recent progress in the modeling of urban systems, we are still missing something extremely important. Suggestions are welcome." We shall see that since Krugman's call for suggestions, we have much improved our understanding of the origin of the Zipf's law, which has forced a great rethinking about the origins of cities—and firms, too.

Firm Sizes

We now look at the firm size distribution. Using US Census data, Axtell (2001) puts firms in "bins'" according to their size, as measured by number of employees, and plots the log of the number of firms within a bin. The result in Figure 3 shows a straight line: again, this is a power law. Here we can even run the regression in "density"—that is, plot the number of firms of size approximately equal to x. If a power law relationship holds, then the density of the firm size distribution is $f(x) = b/x^{\zeta+1}$, so the slope in a log-log plot should be $-(\zeta + 1)$ (because $\ln f(x) = -(\zeta + 1) \ln x$ plus a constant). Impressively, Axtell finds that the exponent $\zeta = 1.059$. This demonstrates a "Zipf's law" for firms.

Figure 3
Log *Frequency* versus log *Size* of US firms (by Number of Employees) for 1997



Source: Axtell (2001).
Notes: Ordinary least squares (OLS) fit gives a slope of 2.06 (s.e. = 0.054; $R^2 = 0.99$). This corresponds to a frequency $f(S) \sim S^{-2.059}$, which is a power law distribution with exponent 1.059. This is very close to an ideal Zipf's law, which would have an exponent $\zeta = 1$.

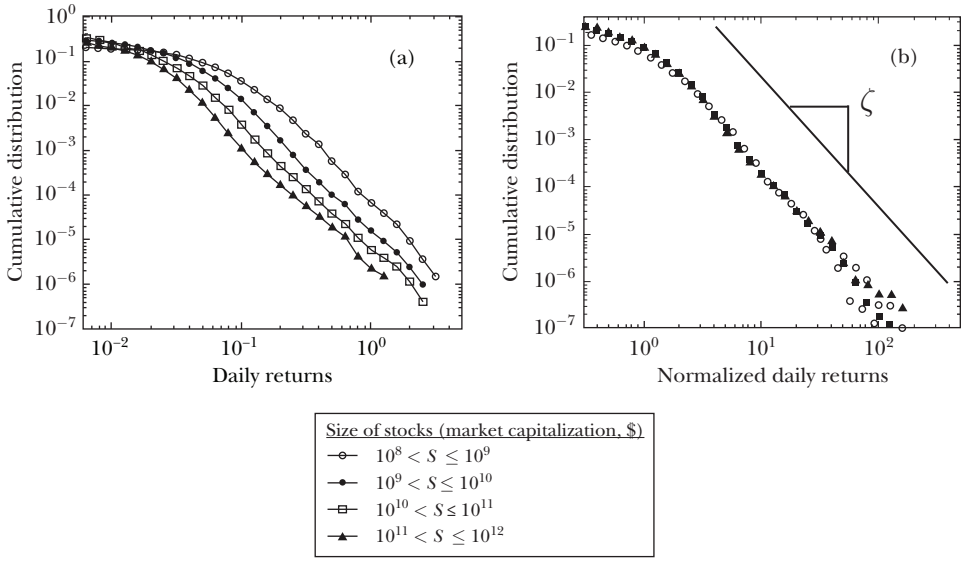
This finding has forced a rethinking of the underpinnings of firms: Most static theories of why firms exist—for example, theories based on economies of scope, fixed costs, elasticity of demand, and the like—would not predict a Zipf's law. Some other type of theory is needed, as we shall soon discuss.

Stock Market Movements

It is well-known that stock market returns are fat-tailed—that is, the probability of finding extreme values is larger than for a Gaussian distribution of the same mean and standard deviation. An energetic movement of physicists, the "econophysicists" (a term coined after the emergence of "geophysicists" and "biophysicists"), has quantified a host of power laws in the stock market. For instance, the size of daily stock market movements are represented in Figure 4. They are consistent with: $\mathrm{P}(|\, r_t \,| > x) = a/x^{\zeta}$ with $\zeta = 3$, the so-called "cubic" law of stock market returns. The left panel of Figure 4 plots the distribution for four different sizes of stocks. The right panel plots the distribution of normalized stock returns, which is calculated as the stock returns divided by their standard deviation: after this normalization, the four different distributions "collapse" onto the same curve. This is a type of "universality"—a term much used in the power law literature (and in physics) which means that different systems behave in the same way, after some rescaling. This cubic law appears to hold for a variety of other international stock markets too (Gopikrishnan et al. 1999).

Likewise, lots of other stock market quantities are distributed according to a power law (Plerou, Gopikrishnan, and Stanley 2005; Kyle and Obizhaeva 2014;

Figure 4
Cumulative Distribution of Daily Stock Market Returns for Different Sizes of Stocks



Source: Plerou et al. (1999).
Notes: The left panel shows the distributions for four different sizes (in terms of market capitalization) of stocks. The right panel shows the returns, normalized by volatility. The slopes $-\zeta$ are close to -3, reflecting the "cubic law" of stock market fluctuations: $P(|r| > x) \sim kx^{-3}$. The horizontal axis displays returns as high as 100 standard deviations.

Bouchaud, Farmer, and Lillo 2009). For instance, the number of trades per day is a power law distributed with exponent of 3, while the number of shares traded per time interval has an exponent of 1.5, and the price impact (that is, the size of the price movement when a large volume of shares is bought or sold) is proportional to the volume to the power of 0.5. Later, I discuss theories explaining those facts.

Again, why do social scientists care? One implication of the cubic law is that there are many more extreme events than would occur if the distribution were Gaussian; for example, if the distribution were the result of a series of events of fixed probability, like flipping coins or rolling dice. More precisely, under the cubic law, the chances of a 10 standard deviation event and a 20 standard deviation event are, respectively, $5^3 = 125$ and $10^3 = 1{,}000$ times less likely than a two standard deviation event, whereas if the distribution of returns was Gaussian, the chances of a 10 and 20 standard deviation event would be 10^{22} and 10^{87} times less likely than a two standard deviation event (much, much less likely). Indeed, in a stock market comprising about 1,000 stocks, a 10 standard deviation event happens in practice just about every day.

Seeking an explanation for these kinds of regularities presses us to rethink the functioning of stock markets—later we shall see theories that exactly explain these exponents in a unified way.

Other Examples of Power Laws

Income and wealth also follow roughly power law distributions, as we have known at least since Pareto (1896), who documented power laws relating to the distribution of income and who was the first to document power laws in economics or in any other area of social science (to the best of my knowledge). The distribution of wealth is more unequal than the distribution of income: this makes sense, because differences in growth rate of wealth across individuals (due to differences in returns or frugality) pile up and add an extra source of inequality. Typically, the Pareto exponent is around 1.5 for wealth and between 1.5 and 3 for income (recall that a lower Pareto exponent means a higher degree of inequality in a distribution). Indeed, power laws and random growth processes are rapidly becoming a central tool to analyze inequalities of income and wealth (Piketty and Zucman 2014; Atkinson, Piketty, and Saez 2011; Benhabib, Bisin, and Zhu 2011; Lucas and Moll 2014; Gabaix, Lasry, Lions, and Moll 2015; Toda and Walsh 2015).

What Causes Power Laws?

Here I sketch the two main mechanisms that generate powers laws: 1) random growth models, which generate a power law and 2) transfer of that power law via matching and optimization (that is, the variable generated by random growth is used as an input begetting another power law in another output variable).

Random Growth

The basic mechanism for generating power laws is proportional random growth (Champernowne 1953; Simon 1955). Suppose that we start with an initial distribution of firms, and they grow and shrink randomly with independent shocks, and they satisfy Gibrat's law (Gibrat 1931), which starts with the assumption that all firms have the same expected growth rate and the same standard deviation of growth rate. However, these basic assumptions do not assure a steady state distribution, because they imply a distribution that over time becomes lognormal with larger and larger variance. However, things change altogether if we add to the model some friction that guarantees the existence of a steady state distribution. Suppose for instance that there is also a lower bound on size, so that a firm size cannot go below a given threshold. Then, as it turns out, the model yields a steady-state distribution, and it is a power law, with some exponent ζ that depends on the details of the growth process.

However, while this mechanism generates a power law, the exponent need not be equal to 1. Why might an exponent of 1 arise?

I give one explanation in Gabaix (1999; for a more thorough review, see Gabaix 2009). Suppose that the size of the assumed friction (the lower bound) becomes very small, and that we have a given exogenous population size to allocate in the system (between the different cities or firms). Then, the exponent ζ becomes 1, rather than any other value. One intuition behind this result is as follows: the exponent cannot be below 1 because then the distribution would have infinite mean

(see footnote 3). Indeed, an exponent just above 1 is the smallest consistent with a finite total population. As the friction becomes very small, the exponent becomes the fattest that is consistent with a finite population.

This insight can explain why we observe lots of Zipf's laws with the exponent ζ equal to 1 in the real economy: because a number of economic variables are well-represented by an underlying pattern of proportional random growth with a small friction and some adding-up constraint for the total size of the system.

Of course, this explanation so far is mechanical; that is, it just points out that a certain process leads to a certain distribution. A social scientist will want to know why these variables exhibit proportional random growth in the first place. Fortunately, once the basic mechanics are clarified, one can suggest good economic reasons as to why they might plausibly exist.

The simplest microfoundation suggested by power laws and Gibrat's law is the following: cities and firms largely exhibit constant returns to scale, perhaps with small deviations from that benchmark, and lots of randomness. In this spirit, many fully economic models for the random growth of cities and firms have been proposed since the 2000s which add more economics to this random growth mechanism (for example, Rossi-Hansberg and Wright 2007; Luttmer 2007). The shocks to this system come from shocks to productivities or amenities. A certain death rate for firms may arise from the processes of creative destruction, and the minimum size can come from a fixed cost that the firm needs to pay to be alive.

As Gibrat's rule of proportionate growth is most naturally consistent with constant return to scale economies, it is an interesting question as to how it fits with the many standard economic theories of firm size that emphasize increasing returns to scale as well as with economies and diseconomies of agglomeration. One possibility is that the effects of increasing returns to scale and the economies and diseconomies of scale are not that big (see the discussion in Rozenfeld, Rybski, Gabaix, and Makse 2009). Another possibility is that these effects are to some extent offset by other large compensating factors, like urban amenities or geography. Yet another possibility is that some of these theories assume that shocks have permanent effects, but some shocks do mean-revert: for example, Japanese cities in large part reverted to their previous sizes after World War II bombing (Davis and Weinstein 2002). Writing richer theories of city size promises to be a fruitful task for future researchers.

Likewise, for the income distribution, the details of the underlying mechanism—say, the extent of luck, the distribution of thrift, and the varying responsiveness to incentives—are very important for a variety of questions, and microfounded models are important. Still, to write sensible theories on this basis one needs to keep in mind the core mechanics of these models, which is proportional random growth leading to power laws.

Matching and Economics of Superstars

Another manifestation of power laws is in the extremely high earnings of top earners in areas of arts, sports, and business. Rosen (1981) suggests a qualitative explanation for this pattern with the "economics of superstars." In Gabaix and

Landier (2008), we present a tractable, calibratable model of this phenomenon along the following lines: Suppose that lots of firms, of different sizes, compete to hire the talents of chief executive officers. In this model, the talent of a chief executive officer (CEO) is given by how much (in percentages) that person is expected to increase the profits of the firm. Competition implements the efficient outcome, which is that the largest firm will be matched with the best CEO in the economy, the second largest firm with the second best CEO, and so on (as in Terviö 2008).[5]

One might think it hopeless to derive a quantitative theory from this starting point, because the distribution of executive talent is very hard to observe. However, we can draw on extreme value theory (which is a branch of probability) to obtain some properties of the tail of the distribution of talent, without knowing the distribution itself. One implication is that, given adjacent chief executive officers in the ordering of talent, the approximate difference in talent between these two CEOs varies like a power law of their rank. The exponent depends on the distribution, but the power law functional form holds for essentially any reasonable distribution (in a way that can be made precise). Given this, in Gabaix and Landier (2008), we work out the pay of CEO number n, who manages a firm of size $S(n)$. We denote by $S(n^*)$ the size of a reference firm, which is the size of the median firm in the Standard and Poor's 500. $D(n^*)$ is a constant that depends on model parameters, like the scarcity of talent. The pay of CEO number n is:

$$w(n) = D(n^*)\, S(n^*)^{1-b} S(n)^b.$$

In this approach, one calibrates $b = 1/3$ (empirically, the exponent tends to be in the [0.3, 0.4] range). For instance, if a firm is eight times bigger than the median firm (so $S(n) = 8\ S(n^*)$), then the CEO of that larger firm earns twice ($8^{1/3}$) as much as the median CEO. But if the size of all firms is multiplied by 8 (so $S(n)$ and $S(n^*)$) are multiplied by 8, the pay of all CEOs is increased by 8.

In this way, the equation creates a "dual scaling" or double power law—because there is a scaling in both average firm size and own firm size. This approach has three implications:

1) Cross-sectional prediction. In a given year, the compensation of a CEO is proportional to the size of the firm to the power of 1/3, $S(n)^{1/3}$, an empirical relationship sometimes called Roberts' (1956) law.

2) Time-series prediction. When the size of all large firms is multiplied by λ (perhaps over a decade), the compensation at all large firms is multiplied by λ. In particular, the pay at the reference firm is proportional to the average market cap of a large firm.

3) Cross-country prediction. Suppose that CEO labor markets are national rather than internationally integrated. For a given firm size S, CEO compensation varies

[5] The microfoundations of matching models and internal organizations are also interesting, and full of power laws (Garicano and Rossi-Hansberg 2006; Geerolf 2015). The availability of microdata makes these detailed microeconomic models even more testable.

across countries with the market capitalization of the reference firm, $S(n^*)^{2/3}$, using the same rank n^* of the reference firm across countries.

It turns out that all three predictions seem to hold empirically since the 1970s. This theory thus points to the increase in firm size as the cause for the increase in CEO pay.

In this way, power laws and extreme value theory are the natural language for drawing quantitative lessons from the "economics of superstars." This formulation explains why very small differences in talent give rise to very large differences in pay: in our calibration in Gabaix and Landier (2008), differences in talent are small and bounded, but differences in pay are unboundedly large. This is what happens when very large firms compete to hire the services of CEOs: small differences of talent, affecting unboundedly large firms, give rise to unboundedly large differences in pay.

The same logic should apply to other markets with superstar characteristics: apartments with a large view of Central Park in New York City, and also top athletes in sports and the price of famous works of art. As far as I know, a systematic quantitative exploration of those issues remains to be done.[6]

This line of thinking leads to a fresh way of thinking about pay–performance sensitivity for chief executive officers. In the classic paper by Jensen and Murphy (1990), they define pay–performance sensitivity as how many dollars the compensation (or wealth) of a CEO changes for a given dollar change in firm value. They find that pay–performance sensitivity is very small: CEOs earn "only" \$3 extra when their firm increases by \$1,000 in value, and so they conclude that corporate governance may not work well. In contrast, in Edmans, Gabaix, and Landier (2009), we propose a different way to think of the benchmark incentives, resting on scaling arguments. Suppose that to motivate the CEO, it is percent/percent incentives, not dollar/dollar incentives that matter: namely, for a 1 percent increase in firm value, the CEO's wealth should increase by k percent, where k is independent of firm size (this relationship is derived from preferences that are multiplicative in effort and consumption).[7] Then, if our earlier expression for the pay of a CEO based on the size of the firm holds, the pay–performance sensitivity in the sense of Jensen and Murphy should decrease as (Firm size)$^{2/3}$. This pattern holds true empirically, as illustrated in Figure 5. Hence, thinking in terms of scaling leads to new thinking about pay–performance sensitivity.

It is thus interesting to study models where assignment and incentives are optimally and jointly determined, a task I carried out with coauthors in Edmans,

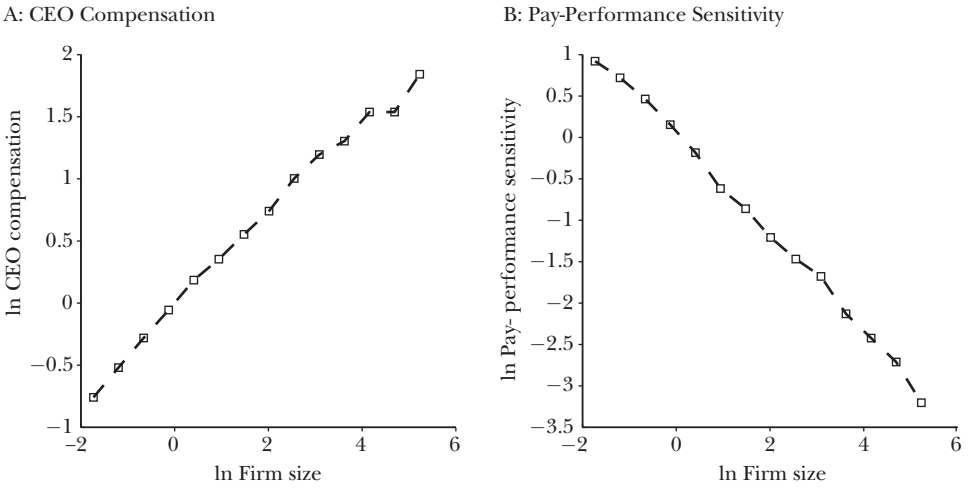
[6] Behrens, Duranton, and Robert-Nicoud (2013) propose a theory of Zipf's law based on matching.

[7] The percent/percent incentive is constant ($d \ln w / dr = k$, where w is CEO pay, and r the firm return), so the Jensen–Murphy pay–performance sensitivity measure is,

$$PPS \coloneqq \frac{dw}{dS} = \frac{dw}{Sdr} = \frac{w}{S}\frac{d\ln w}{dr} = \frac{k' S^b}{S} k = k'' S^{b-1}$$

using $w = k' S^b$ from Roberts' law, and for firm-independent constants k', k''.

Figure 5
CEO Pay and CEO Pay–Performance Sensitivity versus Firm Size



Source: The data and methodology is from Edmans, Gabaix, and Landier (2009), for the years 1994–2008. Years are lumped together by reporting $\ln \frac{w_{it}}{\overline{w}_t}$ (panel A) and $\ln \frac{PPS_{it}}{\overline{PPS}_t}$ (panel B) versus $\ln \frac{S_{it}}{\overline{S}_t}$ (horizontal axis in both panels), where $\overline{x}_t$ indicate the median value of x_{it} in year t.
Notes: Left panel: The CEO compensation is the ex ante one, including Black–Scholes value of options granted. The slope is about 1/3, a reflection of Roberts's law: Pay $\sim$ Size^b with $b \simeq 1/3$. Right panel: The pay–performance sensitivity (PPS) is the Jensen–Murphy measure: by how many dollars does the CEO wealth change, for a given dollar change in firm value. The slope is about $-2/3$, so that PPS $\sim$ Size^{b-1} with $b \simeq 1/3$. The congruence between the scalings is predicted by the Edmans, Gabaix, Landier (2009) model.

Gabaix, and Landier (2009) and Edmans and Gabaix (2011). One result is that pay in the aggregate is all about rewarding talent, not about paying for risk and incentives (which affect pay in the cross-section, not in the aggregate). In the aggregate, the reward to talent fully governs the level of expected pay; incentive issues are quite secondary and simply pin down the form of pay, like what fraction is in fixed or variable pay, not its level. For instance, if some firms are riskier than others, they need to reward their CEOs more (a cross-sectional effect). But if all firms become riskier, the level of pay does not budge (there is no aggregate effect). Hence from that perspective, the rise in pay is all about talent, not incentives.

Optimization and Transfer of Power Laws

Optimization provides a useful way to obtain power laws. For instance, the Allais–Baumol–Tobin rule for the demand for money (which scales as $i^{-1/2}$, where i is the interest rate), is a power law (Allais 1947; Baumol and Tobin 1989). The first scaling relation in economics—and, not coincidentally, the first nontrivial empirical success in economics—may be Hume's thought experiment that doubling the

money supply should lead, after a time, to a doubling of the price level—a basic theory that has stood the test of time.[8]

Power laws have very favorable aggregation properties: taking the sum of two (independent) power law distributions gives another power law distribution. Likewise, multiplying two power laws, taking their max or their min, or a power, gives again a power law distribution. This partly explains the prevalence of power laws: they survive many transformations along with the addition of noise.

Granularity: Aggregate Fluctuations from Microeconomic Shocks

I now turn to an application of power laws: developing a better sense of the origins of aggregate fluctuations in GDP, exports, and the stock market.

Basic Ideas

Where do aggregate fluctuations come from? In Gabaix (2011), I propose that idiosyncratic shocks to firms (or narrowly defined industries) can generate aggregate fluctuations. A priori, many economists would say that this is not quantitatively plausible: there are millions of firms and their idiosyncratic variations should tend to cancel each other out, so the resulting total fluctuations should be very small. However, when the firm size distribution is fat-tailed, this intuition no longer applies, and random shocks to the largest firms can affect total output in a noticeable way.[9]

Empirically, the existence of a power law distribution for firm size suggests that economic activity is indeed very concentrated amongst firms. For instance, di Giovanni and Levchenko (2012) find: "In Korea, the 10 biggest business groups account for 54% of GDP and 51% of total exports. . . . The largest one, Samsung, is responsible for 23% of exports and 14% of GDP." In a setting like this, it seems more plausible that idiosyncratic shocks to firms would affect macroeconomic activity. Likewise, in Japan, the top 10 firms account for 35 percent of exports, and in the United States, the sales of the top 50 firms represent about 25 percent of output (Gabaix 2011). In this view, economic activity is not made of a smooth continuum of firms, but it is made of incompressible "grains" of activities—we call them "firms"—whose fluctuations do not wash out in the aggregate. One plain reason is that some of the firms are very big, and a further reason is that initial shocks can be intensified by a variety of generic amplification mechanisms, such as endogenous changes to hours worked.

Is this granular hypothesis relevant empirically? In Gabaix (2011), I find that idiosyncratic shocks to large firms explain about one-third of GDP fluctuations in

[8] Hume's scaling says $\textit{Price level} = a(\textit{Money supply})^1$, with a proportionality factor a that depends on GDP, a very simple "power law" with exponent of 1.

[9] If there are N firms and a distribution of firms where the central limit theorem applies, the effect of a random shock on total fluctuations should decay as $1/\sqrt{N}$. In a fat-tailed distribution, the standard central limit theorem no longer applies. Instead, a power-law variant holds called the Lévy central limit theorem. In this setting, the effect of random shocks on total GDP fluctuation decays as $1/\ln N$.

the US economy. Di Giovanni, Levchenko, and Mejean (2014) find that they explain over half the fluctuations in France. Further support is given in Foerster, Sarte, and Watson (2011) for industrial production, and in di Giovanni and Levchenko (2012) for exports. The exploration of this theme continues.

This analysis offers two payoffs. First, it may help us to better understand the origins of aggregate fluctuations. Second, these large idiosyncratic shocks may suggest some useful instruments for macroeconomic policy. For instance, Amiti and Weinstein (2013) start from the fact that banking is very concentrated, such that idiosyncratic bank shocks may have strong ripple effects in the aggregate economy; they then seek to quantify these banking channels. They find that idiosyncratic bank shocks can explain 40 percent of aggregate loan and investment fluctuations.

Another implication of granularity is to emphasize the potential importance of networks (Acemoglu, Carvalho, Ozdalar, and Tahbaz-Salehi 2012; Carvalho 2014). Those large firm-level shocks propagate through networks, which create an interesting amplification mechanism and a way to observe the propagation effects. Networks are a particular case of granularity rather than an alternative to it: if all firms had small sales, the central limit theorem would hold and idiosyncratic shocks would all wash out. Networks offer a way to visualize and express the propagation of idiosyncratic firm shocks. Indeed, ongoing research is finding more precise evidence for the explanatory power of this perspective (Kelly, Lustig, and van Nieuwerburgh 2013; Acemoglu, Akcigit, and Kerr 2015).

The Great Moderation: A Granular Post Mortem

This granular perspective also offers a way to understand the time variations in economic volatility. Say that granular or "fundamental" volatility is the volatility that would come only from idiosyncratic sectoral- or firm-level shocks. By construction, when the economy is more diversified, or when the large sectors are in less-volatile industries, fundamental volatility is lower. In Carvalho and Gabaix (2013), we find that fundamental volatility is quite correlated with actual volatility, again consistent with the idea that firm- or industry-specific shocks are an important driver of aggregate fluctuations in the United States and other high-income economies. In addition, policy may dampen or amplify those primitive granular shocks but is not (typically) the primary driver. For instance, in the case of the Great Recession, the primitive shock is a shock to a narrow sector—real estate finance—which was then propagated to the rest of the economy via interesting economic and policy linkages.

This perspective offers an additional narrative for some events of the US economy in recent decades. The US economy experienced what was often called a Great Moderation of lower volatility from the mid-1980s through the mid-2000s, which is often credited at least in part to greater stability of monetary policy. However, from a granularity perspective, the long and large decline of volatility can be traced back to the long and large decline in fundamental volatility at the same period—which came about in part because of the shrinkage of a handful of heavy-manufacturing sectors, whose demise made the economy more diversified. The burst of economic volatility in the 1970s can be attributed in part to the increased importance of a

single sector—the energy sector—which itself can be traced to the rise of oil prices. From this view, the growth in the size of the financial sector is an important determinant of the increase in fundamental volatility—and of actual volatility—in the 2000s. Rather than relying on abstract shocks, a granularity perspective helps us to understand concretely the (proximate) origin of macroeconomic developments.

Linked Volatility of Firms and National Economies

The volatility of the growth rate of firms varies by size, with larger firms tending to have a smaller proportional standard deviation than smaller firms. The volatility of national economies also varies by size, with larger economies tending to have less volatility. Intriguingly, this relationship for firms looks much the same as it does for national economies.

Stanley et al. (1996) study how the volatility of the growth rate of firms changes with size, looking at data for all publicly traded US manufacturing firms between 1975 and 1991. To do this, they calculate the standard deviation $\sigma(S)$ of the growth rate of firms' sales, S, and regress its log against log size. They find an approximately linear relationship, displayed in Figure 6: $\ln \sigma^{\text{firms}}(S) = -\alpha \ln S + \beta$. This means that a firm of size S has volatility proportional to $S^{-\alpha}$ with $\alpha = 0.15$.[10] Lee et al. (1998) conduct the same analysis for fluctuations in the gross domestic product (GDP) of 152 countries for the period 1950–1992 and also find a volatility proportional to $S^{-\alpha'}$ with $\alpha' = 0.15$ (Koren and Tenreyro 2013). These size/volatility relationships, for firms and for countries, are both plotted in Figure 6. The slopes are indeed very similar. This may be a type of "universality." This may be explained by a combination of granularity and power laws: if aggregate fluctuations come from microeconomic shocks, and firm sizes follow Zipf's law, then the identical scaling of Figure 6 should hold true (Gabaix 2011).

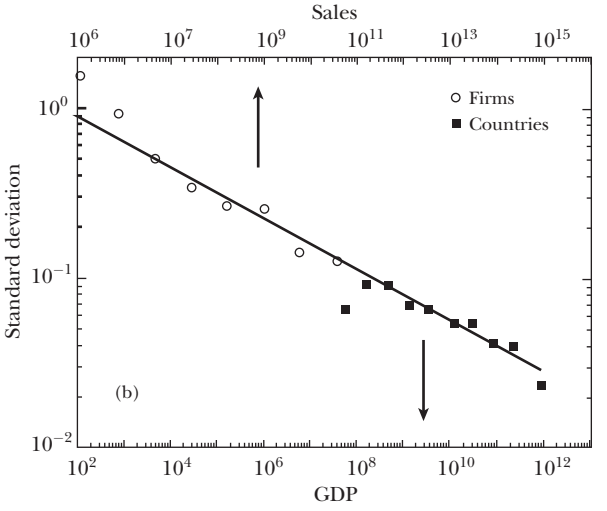
Origins of Stock Market Crashes?

We saw earlier illustrations of power laws in stock market fluctuations. What causes these fluctuations? The power law distribution of firms might actually explain the power law distribution of stock market crashes. In Gabaix, Gopikrishnan, Plerou, and Stanley (2003, 2006), we develop the hypothesis that stock market crashes are due to large financial institutions selling under pressure in illiquid markets (see also Levy and Solomon 1996; Solomon and Richmond 2001). This may account not only for large crashes, but also for the whole distribution of mini-crashes described by the power law.

The power law perspective begins by noting that large institutions are roughly Zipf-distributed, so when they trade, they could have a very large price impact. However, the institutions trade intelligently—and working out the optimal trading strategy of the large institutions gives an explanation for the empirically found

[10] This suggests that large firms are a little more diversified than small firms. However, this diversification effect is weaker than if a firm of size S were composed of S independent units of size 1, which would predict $\alpha = 1/2$. See Riccaboni et al. (2008).

Figure 6
Standard Deviation of the Distribution of Annual Growth Rates



Source: Lee et al. (1998). The firm data are taken from the Compustat for the years 1974–93, the GDP data from Summers and Heston (1991) for the years 1950–92.
Notes: We see that $\sigma(S)$ decays with size S with the same exponent for both countries and firms: $\sigma(S) \sim S^{-\alpha}$, with $\alpha \simeq 1/6$. The size is measured in sales for the companies (top axis) and in GDP for the countries (bottom axis).

power law exponents of trading (an exponent of 3 for returns and 1.5 for volume, as presented earlier). More specifically, large institutions want to moderate their price impact. So, they take more time to execute their large trades. At the optimum, it turns out the price impact they achieve is proportional to the square root of the size of the trade, both in the theory and in the data. Also, in order to moderate total transaction costs, large institutions need to trade proportionally less than smaller ones. In the resulting equilibrium, the distribution of trades is less fat-tailed than the distribution of firms (with an exponent of 1.5), and the distribution of returns is even less fat-tailed (with an exponent of 3)—the factor of 2 coming from the 1/2 exponent of the square root price impact. This theory can be summarized more qualitatively: when large institutions sell under time pressure, they make the market fall and even crash.

One can speculate that this type of mechanism might have been at work in a variety of well-known events, whose origins had one or just a few primitive large traders (which suggests interesting ramifications that are under-researched). The Long Term Capital Management crash in summer 1998 was clearly due to one large fund, with repercussions for large markets (in particular bond markets). The rapid unwinding of very large stock positions by Société Générale after the Kerviel rogue trader scandal caused European stock markets to fall by 6 percent in January 21, 2008, which led the Fed to decrease its rates by 0.75 percentage points. Similarly, it seems the so-called "flash crash" of May 2010 was due to one trader. There is even

tentative evidence that a similar process unwound at the onset of the great stock market crashes of 1929 and 1987 (see the discussion in Gabaix et al. 2006; Kyle and Obizhaeva 2013). This research potentially brings us closer to understanding the origins of stock market movements, with a useful blend of narrative concreteness and general underlying mechanisms leading to clean predictions in the forms of power laws.[11]

Power Laws Outside Economics

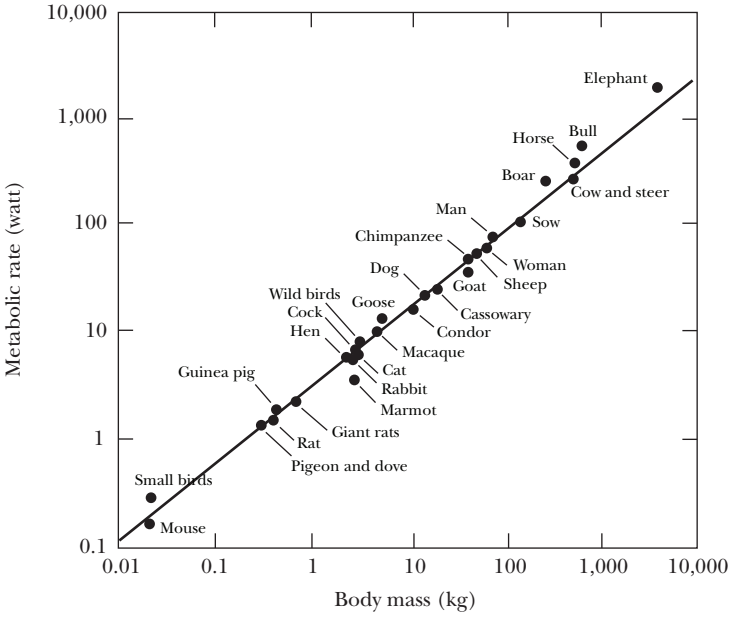
Power laws turn out to be a useful tool for analysis in many areas. For example, natural phenomena like forest fires and rivers have a power law structure (Turcotte 1997). There are concepts here underlying the power law structure that have not (yet) found widespread use in economics. For example, the concept of self-organized criticality considers how a dynamic system can converge to a "critical" state where power law fluctuations occur (for example, a sand pile with many avalanches). The concept of percolation began as the study of how a fluid would filter through a random medium, but has also found applications in other areas like how immunization affects the spread of epidemics. For some applications of these ideas in economics, see Bak, Chen, Scheinkman, and Woodford (1993) and Nirei (2006).

Networks are full of power laws: for instance (probably because of random growth), the popularity of websites, as measured by the number of sites linking to a website, can be represented by a power law (Barabasi and Albert 1999).

Biology is replete with intriguing and seemingly universal relations of the power law type. For instance, the energy that an animal of mass M requires to function is proportional to $M^{3/4}$—rather than M as a simple "constant return to scale" model would predict, as illustrated in Figure 7. West et al. (1997) have proposed the following explanation: If one wants to design an optimal network system to send nutrients to the animal, one designs a fractal system; the resulting efficiency generates the $M^{3/4}$ law. There is an interesting lesson: a priori, lots of things could matter for energy consumption—for example, climate, predator or prey status, thickness of the fur—and they probably do matter to a limited extent. However, in its essence, an animal is best viewed as a network in which nutrients circulate at maximum efficiency. Understanding the power laws forces the researcher to forget, in the first pass, about the details. Likewise, this research shows similar laws for a host of variables, including life expectancy (which scales as $M^{1/4}$). Here a possible interpretation is that the animal is constructed optimally, given engineering constraints that are biologically determined.

[11] Another related perspective on power laws in the stock market concerns potential disasters. Barro and Jin (2011) document a power law distribution of macroeconomic disasters, which may explain abrupt shocks to valuations and indeed many puzzles in finance (Gabaix 2012; Kelly and Jiang 2014). This offers a potentially fruitful direction of research, even if it is still just a hypothesis. See Fu et al. (2005) and Riccaboni et al. (2008).

Figure 7
Metabolic Rate as a Function of Mass across Animals



Source: West, Brown, and Enquist (2000).
Notes: "Metabolic rate" is energy requirement per day. The slope of this log-log graph is 3/4: the metabolic rate of an animal of mass M is proportional to $M^{3/4}$ (Kleiber's law).

Perhaps surprisingly, this type of mechanism, generating power laws from maximum fitness, doesn't seem to have been much studied in economics. For instance, the economy resembles a network with power-law-distributed firms: does this pattern arise from optimality as opposed to randomness? It would be nice to know. Likewise, Zipf's law holds for the usage frequency of words. The simplest explanation is via random growth (as the popularity of words follow a random growth process). However, this law might instead reflect an "optimal" organization of mental categories, perhaps in some tree-like structure? Again, one would like to know.

Conclusion

All economists should become familiar with power laws and the basic mechanisms that generate them because power laws are everywhere. One place to teach power laws is in the macro sequence when discussing models with heterogeneous agents and sectors. The future of power laws as a subject of research looks very healthy: when datasets contain enough variation in some "size"-like factor, such as income or number of employees, power laws seem to appear almost invariably. In addition, power laws can guide the researcher to the essence of a phenomenon.

For instance, consider city size. Lots of things might conceivably be important for city size: specialization, transportation cost, congestion, positive externalities in human capital, and others. The power law approach concludes that while those things exist, they are not the essence of what determines the distribution of city sizes: the essence is random growth with a small friction. To generate the random growth, a judicious mix of the traditional ingredients in the study of cities may be useful, but to orient understanding, one should first think about the essence, and after that about the economic underpinnings.

Many open questions remain about the prevalence and explanation of power laws, and in many of these areas, new data have recently become available. Along with the earlier examples of the distribution of income, wealth, firm size, and city size, here is a sampling of some other questions.

In the study of international trade, the "gravity equation" suggests that the trade flow between any two countries is proportional to the GDP of the two countries (a result which can be derived from a simple model of economies with constant returns to scale) and declines with distance, a finding which seems intuitive. However, the relationship between volume of trade and distance seems to decline with the inverse of distance to the power 1—and there doesn't seem to be any obvious intuitive reason for this particular scale factor. As one possible underlying reason, Chaney (2013) proposes an ingenious model linking this distance to the probability of forming a link in a random growth model of networks, which, under Zipf's law, generates the appropriate coefficient. Similar scaling holds for migration (Levy 2010; Levy and Goldenberg 2014), raising similar questions. For a discussion of power laws in trade, useful starting points include Helpman, Melitz, and Yeaple (2004) and Eaton, Kortum, and Kramarz (2011).

Why is aggregate production in a high-income economy (roughly) Cobb–Douglas with a capital share about 1/3? Jones (2005) generates the functional form but not the particular exponent. Perhaps finding a way to generate this exponent will suggest a deeper understand of the causes of technical progress.

In the study of networks and granularity, how big is the volatility generated from idiosyncratic shocks propagated and amplified in networks?

Is random growth the fundamental origin of power law relationships for the distribution of cities and firms? Or is there potentially some other very different underlying force, like the economics of superstars, or efficiency maximization? Although there is evidence that Gibrat's law seems to roughly hold (that is, the mean and variance of growth rates of a given city or firm is roughly independent of its size, see Ioannides and Overman 2003; Eeckhout 2004), the issue is not settled, as the literature hasn't fully differentiated between permanent shocks and transitory ones, and plain measurement error.

As more of the huge datasets often referred to as "big data" become available, it will be important to characterize and order them. Scaling questions are a natural way to do that and have met with great success in the natural sciences.

A reader seeking a gentle introduction to power law techniques might begin with Gabaix (1999) and then move on to more systematic exposition in Gabaix

(2009), which contains many other pointers. For extreme value theory, Embrechts, Klüppelberg, and Mikosch (1997) is very pedagogical. For networks, Jackson (2008) and Newman (2010) are now standard references. Mantegna and Stanley (2000) contains an accessible introduction to the field of econophysics, while Sornette (2006) contains many interesting physics mechanisms generating scaling.

■ *For excellent research assistance, I thank Chenxi Wang and Jerome Williams, and for comments, I thank David Autor, Chang-Tai Hseih, Chad Jones, Bryan Kelly, Moshe Levy, Timothy Taylor, and David Weinstein.*

References

Acemoglu, Daron, Ufuk Akcigit, and William Kerr. 2015. "Networks and the Macroeconomy: An Empirical Exploration." *NBER Macroeconomics Annual.*

Acemoglu, Daron, Vasco M. Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. 2012. "The Network Origins of Aggregate Fluctuations." *Econometrica* 80(5): 1977–2016.

Allais, Maurice. 1947. *Economie et Intérêt.* Paris: Imprimerie Nationale.

Amiti, Mary, and David Weinstein. 2013. "How Much Do Bank Shocks Affect Investment? Evidence from Matched Bank-Firm Loan Data." NBER Working Paper 18890.

Atkinson, Anthony B, Thomas Piketty, and Emmanuel Saez. 2011. "Top Incomes in the Long Run of History." *Journal of Economic Literature* 49(1): 3–71.

Axtell, Robert L. 2001. "Zipf Distribution of US Firm Sizes." *Science* 293(5536): 1818–20.

Bak, Per, Kan Chen, José Scheinkman, and Michael Woodford. 1993. "Aggregate Fluctuations from Independent Sectoral Shocks: Self-organized Criticality in a Model of Production and Inventory Dynamics." *Ricerche Economiche* 47(1): 3–30.

Barabási, Albert-László, and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286(5439): 509–512.

Barro, Robert J., and Tao Jin. 2011. "On the Size Distribution of Macroeconomic Disasters." *Econometrica* 79(5): 1567–89.

Baumol, William J., and James Tobin. 1989. "The Optimal Cash Balance Proposition: Maurice Allais' Priority." *Journal of Economic Literature* 27(3): 1160–62.

Behrens, Kristian, Gilles Duranton, and Frédéric Robert-Nicoud. 2013. "Productive Cities: Sorting, Selection, and Agglomeration." *Journal of Political Economy* 122(3): 507–553.

Benhabib, Jess, Alberto Bisin, and Shenghao Zhu. 2011. "The Distribution of Wealth and Fiscal Policy in Economies with Finitely Lived Agents." *Econometrica* 79(1): 123–57.

Bouchaud, Jean-Philippe, J. Doyne Farmer, and Fabrizio Lillo. 2009. "How Markets Slowly Digest Changes in Supply and Demand." In *Handbook of Financial Markets: Dynamics and Evolution*, edited by Klaus Reiner Schenk-Hoppe and Thorsten Hens, 57–160. North Holland.

Carvalho, Vasco. 2014. "From Micro to Macro via Production Networks." *Journal of Economic Perspectives* 28(4): 23–48.

Carvalho, Vasco, and Xavier Gabaix. 2013. "The Great Diversification and Its Undoing." *American Economic Review* 103(5): 1697–1727.

Champernowne, D. G. 1953. "A Model of Income Distribution." *Economic Journal* 63(250): 318–51.

Chaney, Thomas. 2013. "The Gravity Equation in International Trade: An Explanation." NBER Working Paper 19285.

Davis, Donald R., and David E. Weinstein. 2002. "Bones, Bombs and Break Points: The Geography of Economic Activity." *American Economic Review* 92(5): 1269–89.

di Giovanni, Julian, and Andrei A. Levchenko. 2012. "Country Size, International Trade, and Aggregate Fluctuations in Granular Economies." *Journal of Political Economy* 120(6): 1083–1132.

di Giovanni, Julian, Andrei Levchenko, and Isabelle Mejean. 2014. "Firms, Destinations, and

Aggregate Fluctuations." *Econometrica* 82(4): 1303–40.

Dittmar, Jeremiah. 2011. "Cities, Markets, and Growth: The Emergence of Zipf's Law." http://www.jeremiahdittmar.com/files/Zipf_Dittmar.pdf.

Eaton, Jonathan, Samuel Kortum, and Francis Kramarz. 2011. "An Anatomy of International Trade: Evidence from French Firms." *Econometrica* 79(5): 1453–98.

Edmans, Alex, and Xavier Gabaix. 2011. "The Effect of Risk on the CEO Market." *Review of Financial Studies* 24(8): 2822–63.

Edmans, Alex, Xavier Gabaix, and Augustin Landier. 2009. "A Multiplicative Model of Optimal CEO Incentives in Market Equilibrium." *Review of Financial Studies* 22(12): 4881–4917.

Eeckhout, Jan. 2004. "Gibrat's Law for (All) Cities." *American Economic Review* 94(5): 1429–51.

Embrechts, Paul, Claudia Klüppelberg, and Thomas Mikosch. 1997. *Modelling Extremal Events for Insurance and Finance.* New York: Springer.

Foerster, Andrew, Pierre-Daniel Sarte, and Mark Watson. 2011. "Sectoral versus Aggregate Shocks: A Structural Factor Analysis of Industrial Production." *Journal of Political Economy* 119(1): 1–38.

Fu, Dongfeng, Fabio Pammolli, Sergey V. Buldyrev, Massimo Riccaboni, Kaushik Matia, Kazuko Yamasaki, and H. Eugene Stanley. 2005. "The Growth of Business Firms: Theoretical Framework and Empirical Evidence." *PNAS* 102(52): 18801–06.

Gabaix, Xavier. 1999. "Zipf's Law for Cities: An Explanation." *Quarterly Journal of Economics* 114(3): 739–67.

Gabaix, Xavier. 2009. "Power Laws in Economics and Finance." *Annual Review of Economics* 1(1): 255–93.

Gabaix, Xavier. 2011. "The Granular Origins of Aggregate Fluctuations." *Econometrica* 79(3): 733–72.

Gabaix, Xavier. 2012. "Variable Rare Disasters: An Exactly Solved Framework for Ten Puzzles in Macro-finance." *Quarterly Journal of Economics* 127(2): 645–700.

Gabaix, Xavier. 2014. "A Sparsity-Based Model of Bounded Rationality." *Quarterly Journal of Economics* 129(4): 1661–1710.

Gabaix, Xavier, Parameswaran Gopikrishnan, Vasiliki Plerou, and H. Eugene Stanley. 2003. "A Theory of Power-Law Distributions in Financial Market Fluctuations." *Nature* 423(6937): 267–70.

Gabaix, Xavier, Parameswaran Gopikrishnan, Vasiliki Plerou, and H. Eugene Stanley. 2006. "Institutional Investors and Stock Market Volatility." *Quarterly Journal of Economics* 121(2): 461–504.

Gabaix, Xavier, and Rustam Ibragimov. 2011. "Rank-1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents." *Journal of Business Economics and Statistics* 29(1): 24–39.

Gabaix, Xavier, and Augustin Landier. 2008. "Why Has CEO Pay Increased So Much?" *Quarterly Journal of Economics* 123(1): 49–100.

Gabaix, Xavier, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll. 2015. "The Dynamics of Inequality." NBER Working Paper 21363.

Garicano, Luis, and Estaban Rossi-Hansberg. 2006. "Organization and Inequality in a Knowledge Economy." *Quarterly Journal of Economics* 121(4): 1383–1435.

Geerolf, François. 2015. "A Static and Microfounded Theory of Zipf's Law for Firms and of the Top Labor Income Distribution." Working Paper, UCLA.

Gibrat, Robert. 1931. *Les inégalités économiques.* Recueil Sirey.

Gopikrishnan, Parameswaran, Vasiliki Plerou, Luis A. Nunes Amaral, Martin Meyer, and H. Eugene Stanley. 1999. "Scaling of the Distribution of Fluctuations of Financial Market Indices." *Physical Review E* 60(5): 5305–16.

Helpman, Elhanan, Marc J. Melitz, and Stephen R. Yeaple. 2004. "Export versus FDI with Heterogeneous Firms." *American Economic Review* 94(1): 300–316.

Ioannides, Yannis M., and Henry G. Overman. 2003. "Zipf's Law for Cities: An Empirical Examination." *Regional Science and Urban Economics* 33(2): 127–37.

Jackson, Matthew O. 2008. *Social and Economic Networks.* Princeton University Press.

Jensen, Michael C., and Kevin J. Murphy. 1990. "Performance Pay and Top-Management Incentives." *Journal of Political Economy* 98(2): 225–65.

Jones, Charles I. 2005. "The Shape of Production Functions and the Direction of Technical Change." *Quarterly Journal of Economics* 120(2): 517–49.

Kelly, Bryan, and Hao Jiang. 2014. "Tail Risk and Asset Prices." *Review of Financial Studies* 27(10): 2841–71.

Kelly, Bryan, Hanno Lustig, and Stijn Van Nieuwerburgh. 2013. "Firm Volatility in Granular Networks." NBER Working Paper 19466.

Koren, Miklós, and Silvana Tenreyro. 2013. "Technological Diversification." *American Economic Review* 103(1): 378–414.

Krugman, Paul. 1996. "Confronting the Mystery of Urban Hierarchy." *Journal of the Japanese and International Economies* 10(4): 399–418.

Kyle, Albert, and Anna Obizhaeva. 2013. "Large Bets and Stock Market Crashes." AFA 2013 San Diego Meetings Paper. Available at SSRN: http://ssrn.com/abstract=2023776.

Kyle, Albert, and Anna Obizhaeva. 2014. "Market Microstructure Invariance: Theory and

Empirical Tests." Available at SSRN: http://ssrn.com/abstract=1687965.

Lee, Youngki, Luís A. Nunes Amaral, David Canning, Martin Meyer, and H. Eugene Stanley. 1998. "Universal Features in the Growth Dynamics of Complex Organizations." *Physical Review Letters* 81(15): 3275–78.

Levy, Moshe. 2010. "Scale-free Human Migration and the Geography of Social Networks." *Physica A: Statistical Mechanics and Its Applications* 389(21): 4913–17.

Levy, Moshe, and Jacob Goldenberg. 2014. "The Gravitational Law of Social Interaction." *Physica A* 393: 418–26.

Levy, Moshe, and Sorin Solomon. 1996. "Power Laws are Logarithmic Boltzmann Laws." *International Journal of Modern Physics C* 7(4): 595–601.

Lucas, Robert E., and Benjamin Moll. 2014. "Knowledge Growth and the Allocation of Time." *Journal of Political Economy* 122(1): 1–51.

Luttmer, Erzo G. J. 2007. "Selection, Growth, and the Size Distribution of Firms." *Quarterly Journal of Economics* 122(3): 1103–44.

Mantegna, Rosario N., and H. Eugene Stanley. 2000. *An Introduction to Econophysics: Correlations and Complexity in Finance.* Cambridge University Press.

Newman, Mark E. J. 2010. *Networks: An Introduction*. Oxford University Press.

Nirei, Makoto. 2006. "Threshold Behavior and Aggregate Fluctuation." *Journal of Economic Theory* 127(1): 309–322.

Pareto, Vilfredo. 1896. *Cours d.économie politique.* Librairie Droz.

Persky, Joseph. 1992. "Retrospectives: Pareto's Law." *Journal of Economic Perspectives* 6(2): 181–92.

Piketty, Thomas, and Gabriel Zucman. 2014. "Capital is Back: Wealth-Income Ratios in Rich Countries, 1700–2010." *Quarterly Journal of Economics* 129(3): 1255–1310.

Plerou, Vasiliki, Parameswaran Gopikrishnan, Luis A. Nunes Amaral, Martin Meyer, and H. Eugene Stanley. 1999. "Scaling of the Distribution of Price Fluctuations of Individual Companies." *Physical Review E* 60(6): 6519–29.

Plerou, Vasiliki, Parameswaran Gopikrishnan, and H. Eugene Stanley. 2005. "Quantifying Fluctuations in Market Liquidity: Analysis of the Bid-Ask Spread." *Physical Review E* 71(4): 046131.

Riccaboni, Massimo, Fabio Pammolli, Sergey V. Buldyrev, Linda Ponta, and H. Eugene Stanley. 2008. "The Size Variance Relationship of Business Firm Growth Rates." *PNAS* 105(50): 19595–19600.

Roberts, David R. 1956. "A General Theory of Executive Compensation Based on Statistically Tested Propositions." *Quarterly Journal of Economics* 70(2): 270–94.

Rosen, Sherwin. 1981. "The Economics of Superstars." *American Economic Review* 71(5): 845–58.

Rossi-Hansberg, Esteban, and Mark L. J. Wright. 2007. "Urban Structure and Growth." *Review of Economic Studies* 74(2): 597–624.

Rozenfeld, Hernán D., Diego Rybski, Xavier Gabaix, and Hernán A. Makse. 2011. "The Area and Population of Cities: New Insights from a Different Perspective on Cities." *American Economic Review* 101(5): 2205–25.

Samuelson, Paul A. 1969. "The Way of an Economist." *International Economic Relations: Proceedings of the Third Congress of the International Economic Association*, pp. 1–11. London: MacMillan.

Simon, Herbert A. 1955. "On a Class of Skew Distribution Functions." *Biometrika* 44(3/4): 425–40.

Solomon, Sorin, and Peter Richmond. 2001. "Power Laws of Wealth, Market Order Volumes and Market Returns." *Physica A: Statistical Mechanics and its Applications* 299(1): 188–97.

Sornette, Didier. 2006. *Critical Phenomena in Natural Sciences.* Springer Science & Business.

Stanley, Michael H. R., Luís A. N. Amaral, Sergey V. Buldyrev, Shlomo Havlin, Heiko Leschhorn, Philipp Maass, Michael A. Salinger, and H. Eugene Stanley. 1996. "Scaling Behaviour in the Growth of Companies." *Nature* 379(6568): 804–06.

Summers, Robert, and Alan Heston. 1991. "The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950–1988." *Quarterly Journal of Economics* 106(2): 327–68.

Terviö, Marko. 2008. "The Difference that CEOs Make: An Assignment Model Approach." *American Economic Review* 98(3): 642–68.

Toda, Alexis, and Kieran Walsh. 2015. "The Double Power Law in Consumption and Implications for Testing Euler Equations." *Journal of Political Economy* 123(5): 1177–1200.

Turcotte, Donald L. 1997. *Fractals and Chaos in Geology and Geophysics.* Cambridge University Press.

US Bureau of the Census. 2012. *Statistical Abstract of the United States: 2012*. http://www.census.gov/compendia/statab/2012edition.html.

West, Geoffrey B., James H. Brown, and Brian J. Enquist. 1997. "A General Model for the Origin of Allometric Scaling Laws in Biology." *Science* 276(5309): 122–26.

West, Geoffrey B., James H. Brown, and Brian J. Enquist. 2000. "The Origin of Universal Scaling Laws in Biology." In *Scaling in Biology*, edited by J. H. Brown, and G. B. West, 87–112. Oxford University Press.

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort.* Cambridge: Addison-Wesley.

Journal of Economic Perspectives—Volume 30, Number 1—Winter 2016—Pages 207–224

Roland Fryer: 2015 John Bates Clark Medalist

Lawrence F. Katz

Roland Fryer has emerged during the last decade as a leading scholar of the US racial divide and as a major figure in the evaluation of education policies to narrow the racial achievement gap and improve the prospects of low-income and minority children. He has been bold and fearless in his willingness to apply rigorous economic theory, to collect new data, and to develop and implement appropriate and compelling empirical strategies (including randomized field experiments) to assess any serious hypothesis that might shed light on racial inequality and that may provide policy tools to improve the academic achievement and long-run outcomes of disadvantaged children. Fryer's work is marked by a creative and entrepreneurial edge that has allowed him to carry out large-scale evaluations and interventions in the context of large school districts like New York, Chicago, Houston, and Dallas, often in the face of political opposition and bureaucratic inertia. His theoretical and empirical work on the "acting white" hypothesis of peer effects provides new insights into the barriers to increasing the educational investments of minorities and the socially excluded. Fryer's research output related to racial inequality, the US racial achievement gap, and the design and evaluation of educational policies make him a worthy recipient of the 2015 John Bates Clark Medal.

He was born in Daytona Beach, Florida. His tumultuous childhood and youth experiences in Florida and Texas have been well-documented in a profile in the *New York Times Magazine* (Dubner 2005). Roland went on to earn his BA from the University of Texas at Arlington in 1998 and his PhD in economics from Pennsylvania State

■ *Lawrence F. Katz is Elisabeth Allison Professor of Economics, Harvard University, and Research Associate, National Bureau of Economic Research, both in Cambridge, Massachusetts. His email address is lkatz@harvard.edu.*

http://dx.doi.org/10.1257/jep.30.1.207 doi=10.1257/jep.30.1.207

Roland Fryer

University in 2002. Fryer was a doctoral fellow and post-doctoral fellow at the University of Chicago and the American Bar Foundation from 2001 to 2003. Steven Levitt and Glenn Loury played formative roles as Roland's early mentors and collaborators. He arrived at Harvard as a Junior Fellow of the Harvard Society of Fellows in 2003, formally joined the Economics Department faculty as an Assistant Professor in 2006, was promoted to tenure in 2007, and currently is the Henry Lee Professor of Economics at Harvard University.

Fryer is an extraordinary applied microeconomist whose work spans labor economics, the economics of education, and the economics of social interactions, while continually returning to the racial divide, one of America's most profound and long-lasting social problems. I will divide this survey of Roland's research into five categories: the racial achievement gap, education policies and reforms, economics of social interactions, the economics of discrimination and anti-discrimination policies, and further topics involving the black–white racial divide. References to Roland's work in this essay will use the numbers from the selected group of his papers given in Table 1.

The Racial Test Score Gap at Different Ages

The existence of a substantial gap in standardized achievement test scores of US black and white children at school ages is well-known, but Roland advanced

Table 1
Selected Research Papers by Roland Fryer

1. "Understanding the Black-White Test Score Gap in the First Two Years of School," (with Steven D. Levitt). 2004. *Review of Economics and Statistics* 86(2): 447–64.
2. "The Causes and Consequences of Distinctively Black Names," (with Steven D. Levitt). 2004. *Quarterly Journal of Economics* 119(3): 767–805.
3. "An Economic Analysis of 'Acting White,'" (with David Austen-Smith). 2005. *Quarterly Journal of Economics* 120(2): 551–83.
4. "Affirmative Action and Its Mythology," (with Glenn C. Loury). 2005. *Journal of Economic Perspectives* 19(3): 147–62.
5. "The Black-White Test Score Gap through Third Grade," (with Steven D. Levitt). 2006. *American Law and Economics Review* 8(2): 249–81.
6. "A Model of Social Interactions and Endogenous Poverty Traps." 2007. *Rationality and Society* 19(3): 335–66.
7. "A Measure of Segregation Based on Social Interactions," (with Federico Echenique). 2007. *Quarterly Journal of Economics* 122(2): 441–85.
8. "Guess Who's Been Coming to Dinner? Trends in Interracial Marriage Over the 20th Century," 2007. *Journal of Economic Perspectives* 21(2): 71–90.
9. "An Economic Analysis of Color-Blind Affirmative Action," (with Glenn Loury and Tolga Yuret). 2008. *Journal of Law, Economics, and Organization* 24(2): 319–55.
10. "A Categorical Model of Cognition and Biased Decision Making," (with Matthew O. Jackson). 2008. *The B.E. Journal of Theoretical Economics* 8(1).
11. "The Changing Consequences of Attending Historically Black Colleges and Universities," (with Michael Greenstone). 2010. *American Economic Journal: Applied Economics* 2(1): 116–48.
12. "An Empirical Analysis of 'Acting White,'" (with Paul Torelli). 2010. *Journal of Public Economics* 94(5–6): 380–96.
13. "Racial Inequality in the 21st Century: The Declining Significance of Discrimination," 2011. *Handbook of Labor Economics*, 4B: 855–971.
14. "Are High Quality Schools Enough to Increase Achievement Among the Poor? Evidence from the Harlem Children's Zone," (with Will Dobbie). 2011. *American Economic Journal: Applied Economics* 3(3): 158–87.
15. "Financial Incentives and Student Achievement: Evidence from Randomized Trials," 2011. *Quarterly Journal of Economics* 126(4): 1755–98.
16. "Hatred and Profits: Under the Hood of the Ku Klux Klan," (with Steven D. Levitt). 2012. *Quarterly Journal of Economics* 127(4): 1883–1925.
17. "The Plight of Mixed-Race Adolescents," (with Lisa Kahn, Steven D. Levitt, and Jorg Spenkuch). 2012. *Review of Economics and Statistics* 94(3): 621–34.
18. "Getting beneath the Veil of Effective Schools: Evidence from New York City," (with Will Dobbie). 2013. *American Economic Journal: Applied Economics* 5(4): 28–60.
19. "Testing for Racial Differences in the Mental Ability of Young Children," (with Steven D. Levitt). 2013. *American Economic Review* 103(2): 981–1005.
20. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools," 2013. *Journal of Labor Economics* 31(2): 373–407.
21. "Measuring Crack Cocaine and Its Impact," (with Paul S. Heaton, Steven D. Levitt, and Kevin M. Murphy). 2013. *Economic Inquiry* 51(3): 1651–81.
22. "Valuing Diversity," (with Glenn C. Loury). 2013. *Journal of Political Economy* 121(4): 747–74.

(continued)

Table 1—Continued

23. "The Impact of Attending a School with High-Achieving Peers: Evidence from New York City Exam Schools," (with Will Dobbie). 2014. *American Economic Journal: Applied Economics* 6(3): 58–75.
24. "The Potential of Urban Boarding Schools for the Poor: Evidence from SEED," (with Vilsa E. Curto). 2014. *Journal of Labor Economics* 32(1): 65–93.
25. "Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments," 2014. *Quarterly Journal of Economics* 129(3): 1355–1407.
26. "The Impact of Voluntary Youth Service on Future Outcomes: Evidence from Teach for America," (with Will Dobbie). 2015. *B.E. Journal of Economic Analysis and Policy* 15(3): 1031–66.
27. "The Medium-Term Impacts of High-Achieving Charter Schools," (with Will Dobbie). 2015. *Journal of Political Economy* 123(5): 985–1037.
28. "Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment," (with Steven D. Levitt, John List, and Sally Sadoff). 2012. NBER Working Paper No. 18237.
29. "Information and Student Achievement: Evidence from a Cellular Phone Experiment," 2013. NBER Working Paper 19113.
30. "Parental Incentives and Early Childhood Achievement: A Field Experiment in Chicago Heights," (with Steven D. Levitt and John A. List). 2015. NBER Working Paper 21477.
31. "Not Too Late: Improving Academic Outcomes for Disadvantaged Youth," (with Philip J. Cook, Kenneth Dodge, George Farkas, Jonathan Guryan, Jens Ludwig, Susan Mayer, Harold Pollack, and Laurence Steinberg), 2015. Working Paper IPR-WP-15-01, Institute for Policy Research, Northwestern University.

the debate by examining the age profile and sources of the test gap from early childhood through high school in a series of highly influential studies. Although a black–white test-score gap is essentially nonexistent in the first year of life, black children fall behind quickly thereafter, and observable background and school variables cannot explain most of the growth of differences in academic achievement between racial groups after kindergarten.

Perhaps the most novel of Fryer's papers on racial test score gaps is a fascinating study with Levitt [19] examining mental ability differences by race for very young children. Fryer and Levitt find no significant black–white test score gap for one year-olds using the Early Childhood Longitudinal Study-Birth Cohort (ECLS-B) data, a nationally representative sample of more than 10,000 children born in 2001. The children are evaluated using the Bayley Scale of Infant Development, which as they explain is based on "exploring objects (e.g., reaching for and holding objects), exploring objects with a purpose (e.g., trying to determine what makes the ringing sound in a bell), babbling expressively, early problem solving (e.g., when a toy is out of reach, using another object as a tool to retrieve the toy), and naming objects." At the ages of 8–12 months, the black–white gap is close to zero. However, the test score gap has become substantial (although much of it is explained by socioeconomic status) by the time kids are two years old. The pattern of almost no initial test score gap followed by a substantial gap by age two suggests a major role of the cumulative effects of different early-age "environments" by race.

In early work, Fryer in collaboration with Levitt [1] shows a substantial widening of black–white test score gaps from the start of kindergarten through the end of first grade using the Early Childhood Longitudinal Survey (ECLS) data. The racial test score gap is largely explained by racial differences in socioeconomic status at the start of schooling, but it expands in the early grades even after allowing interactions of socioeconomic status with age and grade. Fryer and Levitt also find a slower growth in test scores after the start of kindergarten for schools in predominantly black as opposed to predominantly white neighborhoods. The pattern suggests that school quality differences may play a key role in the growing racial test score gap with age, but the extent of racial segregation of US schools and neighborhoods makes it difficult to sort out the role of within- versus between-school factors. A follow-up paper [5] shows a continuing rise in the black–white test score gap through third grade even within schools, suggesting the growth of racial test score gaps with age goes beyond between-school differences in average school quality. Recent work by Bond and Lang (2013) suggests part of the growth in the black–white reading score gap from kindergarten to third grade in the ECLS documented by Fryer and Levitt may not be robust to somewhat arbitrary scaling decisions arising from the ordinal nature of the test scores.

Fryer's comprehensive chapter in the *Handbook of Labor Economics* [13] on "Racial Inequality in the 21st Century" synthesizes and extends his research with Levitt on the racial achievement gap and distills the implications for current racial differences in adult labor market and social outcomes. He presents estimates from ten large datasets including children up to 17 years old and finds that the racial achievement gap is fairly robust across time, samples, and particular assessments used. He documents that racial differences in labor market outcomes today are greatly reduced when one accounts for differences in standardized test scores. Roland infers from these findings that the major challenge for confronting racial inequality in the 21st century is to understand the obstacles undermining the achievement of black and Hispanic children in primary and secondary school. This work is important in focusing the debate on US racial inequality into how schools, parents, and neighborhoods affect the achievement of minority children.

Education Reform I: Student, Teacher, and Parental Financial Incentives

Fryer (2014) described his first major effort to improve school performance in this way in a lecture: "As befits an arrogant economist, my first thought was that this will be easy: We just have to change the incentives. . . . My solution was to propose that we pay them incentives now to reward good school performance. Oh my gosh, I wish someone had warned me. No one told me this was going to be so incredibly unpopular. People were picketing me outside my house saying I would destroy students' love of learning . . ."

The first step was to experiment with short-term financial incentives for students. Fryer [15] implemented and analyzed randomized field experiments in

over 200 urban schools across three cities where treated students were paid for working hard (reading books in Dallas), doing well on interim standardized assessments (New York City), and earning high grades in class (Chicago). Randomization of treatments occurred at the individual student level. Fryer's working hypothesis was that providing students with such financial incentives could improve student achievement if students understand how to productively increase school effort but (in the absence of short-term incentives) lack sufficient motivation to exert optimal effort (perhaps because they lack information or overly discount the future returns to schooling). In contrast, short-term financial incentives would be ineffective if students lack the resources or knowledge to convert effort to measurable achievement. In addition, financial incentives could reduce performance if they undermine "intrinsic motivation."

Although students appeared to be aware of the financial incentives and self-reported that they were motivated by them, the findings for all three cities were of essentially zero mean impact (although some evidence suggests that the incentives of paying per book read in Dallas were successful in improving student achievement for English-speaking students). Fryer concludes that short-term financial incentives are likely of limited effectiveness in improving the school outcomes of disadvantaged children in large US urban districts. The lack of student knowledge of how to improve performance and lack of complementary inputs (like tutoring and peer encouragement) appear important in why such financial incentives are not effective.

Next, Fryer [20] considered incentives for teachers through a set of randomized trials in New York City middle schools and high schools. In contrast to the positive results for some developing countries (for example, Muralidharan and Sundararaman 2011), Fryer found no effect of teacher incentives (in the form of performance bonuses) on educational outcomes. The study suggests the effectiveness of teacher performance incentives in large urban US school districts may be constrained by the relatively modest size of the incentives that appear possible and the complexity of politically feasible incentive programs.

The lack of efficacy of standard teacher incentives in US schools motivated work by Fryer, along with Steve Levitt, John List, and Sally Sadoff, to see if changes in the framing of teacher financial incentives could enhance their effectiveness by harnessing the power of loss aversion [28]. Fryer and his co-authors implemented a randomized control trial in nine schools in Chicago Heights, Illinois, in which teachers were randomly selected for a pay-for-performance program with the timing and framing of the award randomly varied as well. Some teachers were offered traditional end-of-year bonuses for improving student performance ("Gain" treatment), while others were given a lump sum payment at the beginning of the school year and informed that they would have to return some or all of it if their students did not meet performance targets ("Loss" treatment). Teachers in the "Gain" and "Loss" groups with the same performance earned the same final bonus. The striking finding is that the "Loss" treatment substantially and significantly raised student math performance (although not reading performance), and the traditional "Gain"

treatment had no detectable impact. These findings suggest the framing of teacher incentives could be a relevant policy tool.

Having studied incentives for students and for teachers, Fryer along with Levitt and List [30] then examined parental incentives. The set-up is another randomized field experiment in Chicago Heights, with this one offering financial incentives for low-income parents of pre-school children to attend parenting education sessions and to engage in behaviors designed to increase early childhood cognitive and executive-function skills. The financial incentives did increase parent attendance at parenting classes and are associated with initial improvements in early childhood achievement and readiness for school (particularly for Hispanics and as indicated by assessments of noncognitive skills).

Although some of the incentives tested do have positive effects, a main message from Fryer's forays into student, teacher, and parent short-term financial incentives is that they are unlikely by themselves to be key policy levers for dramatically improving the performance of disadvantaged students in traditional US urban public schools. Thus, Roland's efforts soon shifted to examining whether more comprehensive whole-school reforms can narrow the racial achievement gap and improve the longer-run educational and economic outcomes for disadvantaged children, particularly those growing up in high-poverty areas.

Education Reform II: Charter Schools and Comprehensive Whole-School Reforms

Some schools seem clearly more effective than others. But there is always a question as to whether such success is driven (at least to some extent) by the selection of students who might thrive in any environment, or by principals and teachers who have superb but idiosyncratic skills that cannot be transferred to other schools. Fryer's work in this area began with a focus on testing the effectiveness of what appeared to be high-performing charter schools using estimates generated by the random lottery process for admission to oversubscribed charter schools. He then sought to enumerate the policies that characterize the most effective charter schools, and then to test whether such policy combinations can be utilized successfully in regular public schools. His exciting and impressive recent work [25] indicates that policies from charter schools can be moved into poor-performing public schools in Houston, Denver, and Chicago and can generate large math test score improvements.

The Harlem Children's Zone (HCZ), a 97-block area in Harlem that combines "No Excuses" charter schools with a host of neighborhood services (including early childhood education, after-school programs, college guidance, and family support and community health programs), has provided Fryer with a rich testing ground for the role of school-based and neighborhood-based interventions for boosting children to achieve escape velocity from low-income areas. ("No Excuses" doesn't have a precise meaning, but it typically refers to schools that are built on high expectations

for both academic performance and behavior. Such an approach is manifested by a college preparatory curriculum for all students and an emphasis on the school culture, often backed up by a longer school week and by an unambiguous set of rules and punishments.) Fryer and Will Dobbie have exploited the lottery used for admission to the Promise Academy (a charter school that is part of the HCZ), which allows them to examine the effects of the charter school on student outcomes and also to look at differential impacts for students living in and out of the HCZ neighborhoods. In their first paper [14], they examine short-term impacts on student test scores from gaining entry to the Promise Academy and find improvements in math test scores for middle school and elementary school students that are large enough to close the racial achievement gap, as well as substantial gains in reading for elementary school students. A follow-up study [27] finds similar effects on medium-term outcomes including college enrollment, high school graduation, and reductions in risky behavior (teen pregnancies for females and incarceration for males).

The Dobbie–Fryer studies of the HCZ show similar positive impacts for residents and nonresidents of the HCZ indicating high-quality schools alone even without the other HCZ community resources are effective. Also comparisons of lottery losers on the two sides of HCZ residential border reveal no differences in school test scores and medium-term outcomes indicating the HCZ community resources alone do not have much impact on student achievement.

However, a study of the impact of attending SEED schools, urban boarding schools in DC and Baltimore, suggests that greater exposure to standard English outside the classroom can improve test scores. In [24], Fryer and Vilsa Curto consider evidence from the admissions lotteries of the SEED schools, which combine a No Excuses charter school model with a five-day-a-week residential environment and extensive after-school activities and tutoring. The findings indicate large improvements in both math and reading achievement for poor minority children. The math impacts are similar to other high-performing charter schools without a boarding component. But the reading impacts are larger, which suggests that a change in residential environment from inner-city households in neighborhoods where nonstandard English is often spoken to a setting with standard English predominating could be a pathway to improved reading skills.

After providing compelling lottery-based evidence that certain charter schools substantially improve student performance, Fryer has tried to determine the specific policies behind charter school effectiveness and whether such practices can be transferred to other schools. In [18], Fryer and Dobbie collect data on the inner workings of 39 New York City charter schools from interviews and surveys of principals, teachers, and students, along with administrative data. They correlate the school policies and characteristics with estimates of school effectiveness in raising math and reading test scores using lottery-based estimates for 29 of the schools and quasi-experimental matching estimates for the other ten schools. The typical charter school in their sample has only modest impacts on test scores, but the variation across schools is huge. Traditional school input measures—such as class size, spending per pupil, and teacher certification—have little explanatory

power for school effectiveness. But a bundle of five school policies suggested by their in-depth qualitative research—frequent teacher feedback, the use of data to guide instruction, high-dosage tutoring, increased instructional time, and high expectations—are strongly positively correlated with high-performing schools and explain about 45 percent of the cross-school variation in effectiveness.

A concern is that this bundle of five policies that are highly correlated with charter school performance could be just a proxy for idiosyncratically talented principals and teachers, or perhaps for some other set of unmeasured practices. Thus, the next step was to test in large-scale field experiments whether implementing these five school practices in regular public schools generates similar improvements in student performance.

Following intense political negotiations, the public school system in Houston, Texas, was willing to implement this bundle of five best practices from high-performing charter schools in a group of low-performing, traditional public schools. Fryer [25] implemented both a school-level randomized field experiment among 18 low-performing elementary schools as well as quasi-experimental comparisons of Houston public elementary, middle, and high schools getting and not getting the five practices. The study involved major management efforts across a large number of public schools, along with wide-ranging data collection. The findings (covering two years) indicate that injecting these five best practices from charter schools into traditional Houston public schools significantly increased math achievement by 0.15 to 0.18 standard deviations per year (similar to the impacts of high-performing charter schools) but had little effect on reading achievement (as is the case with similar charter schools as well). Fryer also implemented such practices in public schools in Denver and Chicago with similar positive short-run results using convincing quasi-experimental estimates.

These results suggest that charter-school best practices can be used to increase student performance substantially in low-performing traditional public schools. The elementary school intervention was implemented with no additional financial costs, but the secondary school intervention required modest increases in school spending of $1,837 per student (driven by the costs of high-dosage tutoring) similar to the additional costs of high-performing charter schools. Recent work by Abdulkadiroğlu, Angrist, Hull, and Pathak (2014) reinforces Fryer's findings in showing that charter school takeovers of failing public schools in New Orleans and Boston led to similar large improvements in student performance even for nonvolunteer, incumbent groups of students. Also, Fryer's results are supported by the substantial positive impacts of student achievement (as well as on high school graduation and college going) of the reorganization of large New York City public high schools into Small Schools of Choice (typically implementing charter-school best practices) from research by Bloom and Unterman (2014) using lottery-based estimates.

The additional costs of high-dosage tutoring in Fryer's Houston experiment meant it could not be implemented in all the schools. Indeed, Fryer finds substantially larger improvements in student achievement in middle and secondary schools with the high-dosage tutoring versus those without it. These findings motivated an

attempt to test whether the provision of individualized instruction from high-dosage tutoring alone can improve the academic achievement of disadvantaged youth. Fryer and a group of coauthors [31] implemented a large-scale randomized control trial in 12 public high schools in Chicago for ninth and tenth grade disadvantaged males, in which intensive individualized academic instruction—high-dosage math tutoring from Match education (the provider also used in Fryer's Houston experiment)—was provided to a randomly selected individual students. They report striking results, with increases in math achievement scores from high-dosage tutoring of 0.2 standard deviations, large reductions in course failures, and large increases of 0.5 standard deviations in math grades.

The benefits from high-dosage tutoring are interrelated with another concern: Is the supply of talented teachers available to teach in low-performing inner-city schools adequate for expanding successful practices into these schools? Fryer makes the case that individualized high-dosage tutoring allows the use of recent college graduates who don't need teacher training. Another source of talent in inner-city schools is programs such as Teach for America (TFA). Fryer and Dobbie [26] estimate the effect of voluntary youth service as a teacher in TFA using a discontinuity in the TFA application process and a follow-up web survey. They find that marginal TFA participants are 20 percentage points more likely to work in K–12 schools and education more broadly following the end of their TFA service period than those who just missed the TFA selection cut-off. TFA has been criticized for hiring individuals who stay only briefly in education, but these results suggest that service programs such as TFA potentially can be a mechanism to expand the supply of talented teachers and administrators for US public schools.

Fryer's work in the area of school reform continues to expand in scope and depth. For example, Fryer and Dobbie are working with administrative data from Texas to examine long-run impacts of adult earnings from attending the earliest Texas charter schools, using quasi-experimental comparisons to comparable students in regular public schools. Overall, this remarkable and sustained body of research on school-based interventions shows that those based on the best practices of charter schools can substantially improve student performance even for students in failing US urban public schools and even when the interventions start in middle school or later in secondary school. Moreover, the improvements in student performance also appear in some cases to generate longer-term gains through improved high school graduation, higher college enrollment, and reductions in rates of teen birth and criminal activity.

Economics of Social Interactions, Identity, and Acting White

In some minority communities, "acting white" can be an insult meant to convey that a person is turning their back on the minority culture and instead shaping their behavior, appearance, and speech to correspond with expectations of white culture (Fordham and Ogbu 1986). Depending on how these cultural expectations

are defined and enforced, it's possible for a situation to arise in which the accumulation of education may come into conflict with belonging to a cultural group.

Fryer's first major research involved the development of new conceptual frameworks and empirical evidence to understand better the role of "culture" (or identity or cultural capital) in the creation and persistence of racial and ethnic group differences in educational investments and other social outcomes and behaviors. His early work [6] developed an equilibrium model of "cultural capital"—which can be thought of as group-specific investments valuable for future social interactions with a peer or social group. Group-specific investments to facilitate local interactions can in some circumstances run into conflict with investments to increase economic success in broader society. The model provides an explicit foundation and new testable empirical predictions for "acting white" behavior in poor and middle-class black neighborhoods as well as related apparent subculture behaviors for other minority and immigrant group settings. For example, the model highlights that this tension between cultural capital and accumulating education increases as social mobility or geographic mobility become more realistic options for individuals growing up in poor and segregated communities. It also suggests a nonmonotonic relationship between social isolation (and segregation) and the importance of group-specific investments as a signaling device to indicate that an individual will be "sticking around" the neighborhood.

An aversion to "acting white," as well as rejecting those who try too hard in school or accumulate too much education, can arise from what is known as a "two-audience signaling quandary" in the elegant model of Fryer and David Austin-Smith [3]. The two audiences are potential employers and one's peer group, and the signals like educational investment that lead employers to offer high wages can also induce peer group rejection. Without peer pressure, the model looks like a standard (Spence 1973) signaling model in which educational investments are rising monotonically and continuously in academic ability. With peer pressure, "acting white" equilibria can arise where the highest-ability individuals continue to signal their academic ability with high educational investments, but middle-ability youths pool on a common lower education level to get accepted into their peer group. Improvements in outside labor market opportunities encourage more youth to get high education levels and to leave the peer group, but simultaneously can cause those left in the peer group to invest less in education and end up worse off in the mainstream economy.

These models of cultural capital and "acting white" behavior provide a foundation for empirical work on peer effects in school—particularly relationships between student popularity and academic performance—and in showing how what looks like "oppositional" youth cultures can arise as the equilibrium outcome for "rational" actors in a two-audience signaling model. For example, Fryer and Austin-Smith [3] also have used this framework to illuminate the success of the residential-based Job Corps program that provides job training to youth in geographically isolated sites away from neighborhood peer groups as compared to the failure of the JOBSTART program using the same curriculum but operating in a youth's current residential neighborhood (Cave, Bos, Doolittle, and Toussaint 1993).

In complementary empirical work, Fryer has sought to assess how racial differences in culture (or identity) and in peer group interactions may affect youth educational and social behavior. Fryer and Paul Torelli [12] produced a fascinating empirical study using rich data on friendship networks for high school students from the National Longitudinal Survey of Adolescent Health (AdHealth). They uncover noticeable racial differences in the relationship of student popularity and grades, especially in schools with greater interracial contact. The findings are consistent with Fryer's two-audience signaling model of "acting white" with a less positive impact of grades on popularity for blacks in schools with more interracial contact. This work uses improved measures of friendship networks and on the popularity and racial isolation (segregation) of one's friends based on Fryer's [7] own rigorous work, discussed later, in creating a new index to measure the extent of segregation at the individual level.

In yet another approach to looking at effects of peer groups, Fryer and Dobbie [23] take advantage of the fact that New York City has several elite "examination" high schools, where admission is determined by being above a certain threshold point on an entrance test. As a result, a regression-discontinuity strategy can compare the longer-term experience of students with very similar test scores, some of whom were just above the threshold and thus admitted to a school with a different peer group, and those who were just below the threshold. The authors show that being admitted to an elite examination school leads to a substantial difference in average peer quality; however, they find no effect of admission to an elite high school on college attendance or graduation rates for marginal admits. Of course, the marginal admits may go from getting extra resources and attention as a top student at a non-elite school to being lower-tier students getting less attention at the elite examination schools.

Might an information-based intervention overcome what can be negative peer influences on educational attainment? In a randomized field experiment in Oklahoma City Public Schools, Fryer [29] provided daily information about the link between human capital and future outcomes, via messages to freely provided cellular phones. The information intervention changed students' *reported* beliefs about links between education and life outcomes and reports of school effort. But the treatment had no detectable impacts on school attendance, behavioral problems, or test scores. The findings suggest that disadvantaged students do not know which strategies of studying and learning are needed to translate their beliefs about gains from education into measurable improvements in their own level of education—a finding which echoes the results of Fryer's [15] experiments on student financial incentives mentioned earlier.

Other work has tested and confirmed Fryer's "acting white" hypothesis based on differential peer pressure influences across settings. In a recent example, Bursztyn and Jensen (2015), in a randomized control trial for disadvantaged eleventh grade students in Los Angeles public schools, provide free access to an SAT prep course and randomize whether the decision to take-up the offer is private or made public to one's classroom peers. They find that student take-up is

much lower in the public than private treatment in non-honors classes, but find no difference in take-up by treatment in honors classes. The same pattern holds for honors students across different classroom environments.

The role of racial identity and cultural capital seems to matter for other social behaviors, too. Of particular note is Fryer's work with Levitt [2] examining changes in patterns of first names of black and white children. Based on individual-level California birth certificate data, blacks and whites chose relatively similar first names for their children in the 1960s. But starting in the 1970s that pattern changed substantially, with blacks (particularly those living in racially isolated neighborhoods) adopting increasingly distinctive names. Fryer and Levitt convincingly argue that the rise of the Black Power movement in the late 1960s and the 1970s influenced how blacks perceived their identities and that first names of children provide a useful window for measuring racial identity. They further document that first names increasingly provide a strong signal of socioeconomic background for US blacks in a way that was not previously the case for those born prior to the 1970s.

Economics of Discrimination, Affirmative Action, and Anti-Discrimination Policy

Studies in social psychology indicate that individuals typically process information with the aid of categories. Fryer and Matthew Jackson [10] use this idea to explore the psychological foundations for racial discrimination. In their model, specific biases emerge from a combination of optimal categorization with a fixed number of categories. Types of experiences and objects that are less frequent in the population tend to be more coarsely categorized and lumped together. Decision-makers make less accurate predictions when confronted with such objects. This can lead to discrimination against minority groups that looks like what is called "statistical discrimination"—which refers to treating all members of a minority group as part of one category and not making finer distinctions—even with no malevolent "taste" for discrimination.

The controversies concerning affirmative action often do not reflect the fact that such policies can be implemented in quite different ways. For example, an affirmative action policy can be a "sighted" one that takes race explicitly into account, or an "unsighted" or "color-blind" policy that is based on a factor that will have the effect of advantaging some members of a group without explicitly taking race into account. Examples of color-blind affirmative action policies for college admission include taking into account a student's socioeconomic status (family income or parental education) and guaranteeing admissions to flagship state universities to students in the top 10 percent of their public high school class. The correlations of socioeconomic status with race and the substantial racial segregation of US neighborhoods and public high schools mean such color-blind policies can impact the racial mix of students admitted to a college. Affirmative action policies can also be implemented either in the form of development assistance, like preferential access

to training or schools, or in the form of job placement advantages, like a goal or quota being set for hiring. Such differences need to be taken into account to accurately evaluate the proper mix and efficacy of affirmative action policies.

In expositing the economics and consequences of affirmative action, Fryer has done first-rate work with Glenn Loury in this journal [4] showing both empirically and conceptually the inefficiencies of "color-blind" affirmative action for increasing racial diversity in university admissions or hiring [9]. Fryer and Loury show that the use of color-blind affirmative action policies to try to maintain a given level of racial diversity leads to distortions in college admission rules for all students (minority and nonminority) and can also cause negative feedback into the investments made before college by high school students by suboptimally changing incentives for taking harder courses, getting involved in extracurricular activities, and improving skills evaluated on standardized tests.

In their foundational theoretical piece on affirmative action policies entitled "Valuing Diversity," Fryer and Loury [22] develop a model of a population of agents belonging to distinct social groups who invest in human capital, and then compete for assignments that give them an opportunity to use their skills. One group is disadvantaged, and policies to enhance opportunity for the agents in that group are considered. Relative to the existing literature on affirmative action, Fryer and Loury provide a more rigorous analysis in the tradition of optimal tax theory and provide close attention to the key distinctions like whether such programs are sighted versus unsighted, or whether they are applied at the stage of skill acquisition or the stage of hiring. Their model implies that optimal sighted affirmative action policies should take place at the hiring stage, but optimal color-blind policies may include interventions at the education stage.

Further Topics: Segregation, the Crack Epidemic, Interracial Marriage, HBCUs, and the KKK

Common empirical measures of segregation such as the "dissimilarity index" or "isolation index" have typically been based on group statistics such as the distribution of shares of different groups (defined by race, sex, or ethnicity) by fixed geographic units such as a Census tract or zip code. Such measures are inevitably vulnerable to the concern that different arbitrary partitions of space could lead to very different results. In contrast, Fryer and Federico Echenique [7] develop a new measure, the Spectral Segregation Index, that can be disaggregated to the individual level. The Spectral Segregation Index is based on the premise that an individual is more segregated the more segregated are the agents with whom he or she interacts. Fryer and Echenique show how three reasonable properties for such an index—monotonicity, homogeneity, and linearity—generate the Spectral Segregation Index. In practice, the Spectral Segregation Index looks a lot like the familiar isolation index for measuring the extent of geographic racial segregation. However, because it is calculated from the individual level it can also

reveal the determinants of racial patterns of friendship networks in high schools and a wide range of other social interactions. The Spectral Segregation Index was a concept in advance of the data available when it was developed a decade ago. However, it is now possible to do individual-level measures of segregation on geocoded microdata from the historical US Censuses of Population up to 1940 and with new big online datasets from social networks and mobile apps that track locations of consumer purchases and social interactions. The value of the Spectral Segregation Index for individual-level measure of segregation was demonstrated in Fryer's [12] work on the relationships between grades and popularity by race and school demographics discussed earlier.

Although it is commonly believed that the "crack epidemic" of the 1980s and early 1990s had strong negative effects on black youth and families, detailed empirical studies were hampered by the absence of a direct measure of the prevalence of crack cocaine. However, Fryer [21] produced impressive work with Paul Heaton, Levitt, and Kevin M. Murphy on measuring the prevalence of crack cocaine by city and over time through the construction of an index based on a range of indirect proxies (cocaine arrests, cocaine-related emergency room visits, cocaine-induced drug deaths, crack mentions in newspapers, and DEA drug busts). Their crack index reproduces many of the spatial and temporal patterns described in ethnographic and popular accounts of the crack epidemic: for example, the rise in the crack index in certain cities fits with rising black youth homicides and poor birth outcomes in the late 1980s and early 1990s.

The United States has a long history of legal restrictions on interracial marriage, which evolved across different states at different times. In this journal, Fryer [8] presented a fascinating historical analysis and descriptive work on marriage patterns by race and inter-racial marriage trends from 1880 to 2000 using microdata from the US population censuses. Fryer also explores the explanatory power of a Becker-style marriage model with some racial discriminatory tastes for explaining the patterns in the data. Fryer along with Steven Levitt, Lisa Kahn, and Jorg Spenkuch [17] has empirically examined "The Plight of Mixed Race Adolescents," documenting that mixed-race children have economic outcomes between blacks and whites, but that mixed-race kids have higher levels of risky and problem behaviors as adolescents. Fryer and his coauthors show that a model in the style of Roy (1951) focused on peer interactions can help explain this pattern with mixed-race kids having no predetermined peer group and choosing more risky behaviors to gain acceptance.

Until the 1960s, the "historically black colleges and universities" (HBCUs) were practically the only institutions of higher learning open to many blacks in the United States. However, the role of these institutions has changed substantially since about 1970. Fryer and Michael Greenstone [11] have produced a careful empirical study of the changing role of the historically black institutions in the higher education of US blacks. Using nationally representative data from the 1970s and the 1990s, Fryer and Greenstone find that in the 1970s HBCU matriculation was associated with higher wages and an increased probability of college graduation for blacks, relative to attending a traditionally white institution.

By the 1990s, however, there is a wage penalty for blacks attending the HBCUs, resulting from a 20 percent decline in the relative wages of HBCU graduates between the two decades. They explore a range of explanations for these patterns, and also consider the satisfaction with their educations of blacks attending both historically black and traditionally white institutions. Fryer and Greenstone's data find modest support for the possibility that the relative decline in wages for graduates of historically black institutions is at least partially due to improvements in the effectiveness of traditionally white institutions at educating blacks, rather than declines in the "quality" of the historically black colleges and universities.

Finally, one of the most infamous and grisly racist organizations in US history is the Ku Klux Klan, which claimed millions of members during its heyday in the mid-1920s. Fryer and Levitt [16] analyze the 1920s Klan: who joined it and also its social and political impact. They utilize a wide range of newly discovered data sources: information from Klan membership rolls, applications, and robe-order forms; an internal audit of the Klan by the accounting firm Ernst and Ernst; and a census that the Klan conducted after an internal scandal. Combining these sources with data from the 1920 and 1930 US population censuses, they find that individuals who joined the Klan were better educated and more likely to hold professional jobs than the typical American. During this time period, they uncover few tangible social or political impacts of the Klan and little evidence that the Klan had a large effect on black or foreign-born residential mobility, or on lynching patterns. Fryer and Levitt conclude that the 1920s Klan, at least outside the Deep South, is best described as a social organization built through a wildly successful pyramid scheme in which individuals at the top of the Klan got rich by charging fees and selling robes to individual members, with the entire process fueled by an army of highly incentivized sales agents selling hatred, religious intolerance, and fraternity in a time and place where there appears to have been much "demand" for such a "product."

Conclusion

As this review of his work illustrated, Roland Fryer has shown an extraordinary willingness and ability to master tools from many disciplines to use the most appropriate scientific methodology available to tackle research topics that help illuminate racial inequality and policies to address it. For example, he has designed, implemented, and analyzed large-scale field experiments in low-income urban schools and neighborhoods. Much of this work is done through EdLabs, the Education Innovation Laboratory at Harvard University, which Roland founded in 2008 and where he continues as its director. He has employed a wide variety of other approaches as well, ranging from historical archival research to alternative credible identification strategies including regression-discontinuity designs. He also has done pathbreaking theoretical work on the economics of affirmative action, models of peer effects, and the measurement of segregation.

I have had the pleasure of being able to interact with and learn from Roland on a regular basis since he arrived at Harvard in 2003. Roland combines much good humor and geniality with a seriousness of purpose and laser-beam focus on data, research methods, and theory. For an easily accessible taste of Roland's intellectual approach and persona, interested readers can watch his 2014 Henry and David Bryna Lecture at the National Academy of Sciences on the subject, "21st Century Inequality: Does Discrimination Still Matter?" Video and slides from the lecture are available at http://sites.nationalacademies.org/DBASSE/DBASSE_088044, and an edited version of the talk is published as Fryer (2014). I look forward to continuing to learn from Roland's ongoing projects, including short-run and long-run impacts of school and neighborhood interventions, as well as the measurement of potential racial biases in policing. His innovative and creative research contributions are sure to continue to deepen our understanding of the sources, magnitude, and persistence of the US racial divide and of broader issues related to social and economic inequality.

References

Abdulkadiroğlu, Atila, Joshua D. Angrist, Peter D. Hull, and Parag A. Pathak. 2014. "Charters without Lotteries: Testing Takeovers in New Orleans and Boston." NBER Working Paper 20792, December.

Bond, Timothy N., and Kevin Lang. 2013. "The Evolution of the Black-White Test Score Gap in Grade K-3: The Fragility of Results." *Review of Economics and Statistics* 95(5): 1468–79.

Bloom, Howard S., and Rebecca Unterman. 2014. "Can Small High Schools of Choice Improve Educational Prospects for Disadvantaged Students?" *Journal of Policy Analysis and Management* 33(2): 290–319.

Bursztyn, Leonardo, and Robert Jensen. 2015. "How Does Peer Pressure Affect Educational Investments?" *Quarterly Journal of Economics* 130(3): 1329–67.

Cave, George, Hans Bos, Fred Doolittle, and Cyril Toussaint. 1993. *JOBSTART: Final Report on a Program for School Dropouts*. October. MDRC, Manpower Demonstration Research Corporation. http://www.mdrc.org/sites/default/files/full_416.pdf.

Dubner, Stephen J. 2005. "Toward a Unified Theory of Black America." *New York Times Magazine*, March 20.

Fordham, Signithia, and John Ogbu. 1986. "Black Students' School Success: Coping with the Burden of 'Acting White.'" *Urban Review* 18(3): 176–206.

Fryer, Roland. 2014. "21st Century Inequality: The Declining Significance of Discrimination." *Issues in Science and Technology* 31(1).

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119(1): 39–77.

Roy, A. D. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* (New Series) 3(2): 135–46.

Spence, Michael. 1973. "Job Market Signaling." *Quarterly Journal of Economics* 87(3): 355–74.

Journal of Economic Perspectives—Volume 30, Number 1—Winter 2016—Pages 225–231

Retrospectives
What Did the Ancient Greeks Mean by *Oikonomia*?

Dotan Leshem

This feature addresses the history of economic terms and ideas. The hope is to deepen the workaday dialogue of economists, while perhaps also casting new light on ongoing questions. If you have suggestions for future topics or authors, please write to Joseph Persky of the University of Illinois at Chicago at jpersky@uic.edu.

Introduction

Nearly every economist has at some point in the standard coursework been exposed to a brief explanation that the origin of the word "economy" can be traced back to the Greek word *oikonomia* (οἰκονομία), which in turn is composed of two words: *oikos*, which is usually translated as "household"; and *nemein,* which is best translated as "management and dispensation." Thus, the cursory story usually goes, the term *oikonomia* referred to "household management" and while this was in some loose way linked to the idea of budgeting, it has little or no relevance to contemporary economics.

This article introduces in more detail what the ancient Greek philosophers meant by "oikonomia." It begins with a short history of the word. It then explores some of the key elements of oikonomia, while offering some comparisons and contrasts with modern economic thought. For example, both Ancient Greek oikonomia and contemporary economics study human behavior as a

■ *Dotan Leshem is Senior Lecturer, School of Political Science, University of Haifa, Haifa, Israel. His email address is dotanleshem@yahoo.com.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
http://dx.doi.org/10.1257/jep.30.1.225

doi=10.1257/jep.30.1.225

relationship between ends and means which have alternative uses. However, while both approaches hold that the rationality of any economic action is dependent on the frugal use of means, contemporary economics is largely neutral between ends, while in ancient economic theory, an action is considered economically rational only when taken towards a praiseworthy end. Moreover, the ancient philosophers had a distinct view of what constituted such an end—specifically, acting as a philosopher or as an active participant in the life of the city-state.

In this way, the most striking difference between ancient oikonomia and contemporary economics is their relationship to ethics. Contemporary economics is "fundamentally distinct from ethics" (Robbins 1935, p. 135), and its theory "is in principle independent of any particular ethical position" (Friedman 1953, p. 4). In addition, contemporary economists typically hold that the natural situation for humans is to live in a world in which means are scarce. On the contrary, the ancient Greek writers on oikonomia believed that humans live in a world of natural abundance that is sufficient for what people need for subsistence. From their perspective, the main task of economic rationality is to advance the good life as they understood it, which means support for philosophy, for involvement in public life, and also for not giving in to what they viewed as the unnatural urge to pursue economic goals or luxuries for their own sake. The oikonomia literature was rooted in the society of its time. It focused on well-to-do, land-owning male citizens, and it included an unthinking acceptance of slavery as well as archaic and demeaning attitudes toward women. However, the discussions in the oikonomia literature concerning how to manage slaves offer some embryonic examples of discussions about how to provide incentives for labor; while the figure of the matron, more than any other figure in the ancient Greek oikonomia literature, shares traits with the modern *homo economicus*. That oikonomia is so rooted in ethical judgments raises questions about whether or in what ways modern economics should be linked to a more explicit consideration of what constitutes a good life.

History of the Word "Oikonomia"

In ancient Greece, the "oikos" in oikonomia referred to a household not just in the sense of a family consumption unit, but more in the sense of an estate. An oikos was also a manufacturing unit that supplied many of its own needs and, in many cases, included slaves along with the nuclear family. Although oikos management was thoroughly dealt with in texts from the Archaic period (approximately the 8th to 6th centuries BCE), the word "oikonomia" hardly appears in these texts. For example, Hesiod's *Work and Days* (circa 700 BCE) is dedicated to the management of the oikos and is full of advice about agricultural production, however, in this 800-line didactic poem, the term "oikonomia" does not arise. It seems as if the tacit assumption in writings during this time was that all of life that mattered took place within the bounds of one's oikos. Thus, it was not necessary to offer a separate discussion of economic matters under the subject matter of oikonomia, nor was it

necessary to distinguish between the economic and the political sphere. One exception is the earliest appearance of the root word for "oikonomia," which is found in a poem by Phocilides (6th century BCE) that reflects the misogynic spirit of texts from this age. The poet classifies women by comparing them to different kinds of animals and advises his friends to marry a "bee-like" wife, because she is a "good oikonomos who knows how to work." "Oikonomos" is usually translated as "steward."

But as the affairs of the city-state (the *polis*, hence politics) became more central in the lives of the gentry, the term "oikonomia" came into common use. This period is often called the "classical age" of ancient Greece starting around 500 BCE, which is the time of the rise of the Athenian democracy and Socratic philosophy. The word "oikonomia" was used in speeches before the members of the Athenian court: for example, Socrates (who was born around 470 BCE and took hemlock in 399 BCE) used the term at his trial when claiming that he neglected what most men care for, such as oikonomia (Plato's "Apology": 36b); and Lysias (born circa 459 BCE) recalls in his speech *On the Murder of Eratosthenes* (1.7) that the wife of Euphiletos was a clever and frugal oikonomos. The term "oikonomia" is also found in plays from that period, as in the tragedy *Electra* (190) by Sophocles (born circa 497 BCE) when Electra murmurs that she had to serve as an oikonomos of her father's house after his murder as if she were a despised slave. Electra's comment demonstrates in a backhanded way the improvement in material well-being that many Greek *oikoi*, or estates, experienced during those years, as well as the relative rise in the social status of certain female citizens who were able to dispense some of their duties to slaves.

During this time period around the fifth century BCE, as oikonomia was talked about in the political sphere and put on display in the theaters, it also became a subject of philosophical reflection. Antisthenes (born circa 445 BCE), a philosopher who was the founder of Cynicism and a companion of Socrates, composed the first of these treatises—at least according to Diogenes Laertius (6.16) writing in the third century CE—but that work did not survive. He was soon followed by Xenophon (born circa 430 BCE) and his book *Oikonomikos* (meaning "one who knows economics"), and a generation later by Aristotle (born circa 384 BCE), who wrote a book on oikonomia (Diogenes Laeritus II: 12), of which only two fragments survived. Oikonomia literature was not limited to classical Greece. In the next 500 years of Greek-speaking antiquity—that is, from 332 BCE to roughly 200 CE—all major schools of Greek philosophical thought composed texts dedicated to household management (see references for sources on these authors). More specifically, students of Aristotle composed three texts on economics, which were ascribed to Aristotle himself (conventionally, we refer to these as written by Pseudo Aristotle). By the first century BCE, Philodemus of Gadara ascribed the first of these works to Theophrastus who replaced Aristotle as the head of the school.[1] A summary of Stoic and "Peripatetic" (meaning "Aristotelian") economic thought by Arius Didymus was

[1] For a discussion of the influence of this first book (that is ascribed to Theophrastus) on economic thought and economic policy in the Middle Ages, Renaissance, and early modernity, see Baloglou (2009).

saved, as well as treatises dedicated to economics by members of the Pythagorean (Callicratidas),[2] Stoic (Hierocles), and Epicurean (Philodemus) schools.[3]

By and large, these texts were addressing male citizens who headed well-to-do households and adhered to the values of the landed gentry. This meant that, on top of uncritical acceptance of enslavement and the subjection of women already mentioned, a valorization of self-sufficiency of the household (autarky) and a degree of scorn about market trading infused these works (Leshem 2014a). Such antagonism to the marketplace reflected an opposition to the social classes who were selling goods and labor in the market. As a result, the oikonomia literature is dealing only with social and economic activities that took place in the household (and when doing so, it was typically prescriptive rather than descriptive). Although the share of typical economic activities happening within estates was much greater in the time of ancient Greece than it is today, this focus falls far short of covering the whole of the economy and society of ancient Greece.

Modern academics sometimes write about "economic imperialism," by which they mean to praise or curse the incursions of economists into neighboring fields like history, psychology, political science, sociology, philosophy, and others. The ancient Greek world witnessed its own "economic imperialism" of terminology, as "oikonomia" became a loanword applying to nearly every sphere of life. As I describe in Leshem (2013b), whatever people did, wherever they turned, they were seen as economizing. Both bodily functions and ethical choices were conceived as "economized"—that is, seen as rationally managed; the term political economy appears in reference to Ptolemaic Egypt; and even the cosmos itself was conceived by the Stoic philosophers as rationally economized by Nature. "Oikonomia" was also used as a term denoting the rational management of resources in political theory, military strategy, law, finance, medicine, literary criticism, architecture, music, history, and rhetoric.[4] It is uncertain what caused the rise in the popularity of the word "oikonomia" in the Hellenic and Roman Empires. Recounting other episodes of economic imperialism in the history of the west, I suggest in Leshem (2013b) that it was a byproduct of both the spirit of political expansion that was a hallmark of both Hellenistic and Roman Empires and of a contemporary ideology that sought to identify rational design in nature and culture.

[2] Several texts on the subject of oikos management are ascribed to female members of the Pythagorean school —Theano, Perictione I, Phyntis, Myia, Aesara—but only fragments of these texts have been saved (Waithe 1987, pp. 61, 65, 72–73).

[3] Beside the above-mentioned texts, Albert Augustus Trever (1916, pp. 128–29) refers to texts composed by Xenocrates, Theophrastus, Metrodorus of Lampsacus, and Dio Chrysostom that are only mentioned by other ancient writers, or of which only a few fragments survived. Another genre that can shed light on the study of oikonomia is "on marriage" (for discussion, see Natali 1995).

[4] In the encyclopedic research of Reumann (1957) into the philological history of "oikonomia," he grouped its uses into four categories: i) oikonomia as the management of the oikos (153–205); ii) oikonomia in the political sphere (206–305); iii) oikonomia in nearly every art and science where it usually means the rational use of the field resource (306–390); and iv) oikonomia of the cosmos (391–486).

Ends and Means

Lionel Robbins (1935, p. 16) offered a definition of economics familiar to modern economists: "Economics is the science which studies human behavior as a relationship between ends and scarce means which have alternative uses." In contrast, Xenophon offers this definition in the earliest text of oikonomia that has survived, which was highly influential for several hundred years: "The name of a branch of theoretical knowledge, and this knowledge appeared to be that by which men can increase oikos, and an oikos appeared to be identical with the total of one's property, and we said that property is that which is useful for life, and useful things turned out to be all those things that one knows how to use" (*Oeconomicus* [hereafter "Ec."] 6:4). This definition obviously suffers in translation, but in Leshem (2013c), I address the task of explicating it in the context of the ancient literature on oikonomia. Robbins (1935, p. 24) states that "economics is entirely neutral between ends." In contrast, ancient economics was deeply concerned with ends as such, and in the selection between possible ends. In addition, ancient economics was a science that studied human behavior as a relationship between ends and *abundant* means, which have alternative uses.

In the writings of the ancient Greeks, the life of the head of the household—the *oikodesptes* who was the addressee of these texts—was conducted in three dimensions: the spiritual realm of philosophy, the heroic realm of politics, and the economic realm. The role of the economic dimension was to secure the means necessary for existence and to generate a surplus that sustained the two other dimensions that were deemed worthy of man. This could be done in two ways: either by increasing production or by moderating consumption.

These two options for securing life's necessities and generating surplus can also be derived from Xenophon's discussion. Xenophon portrays philosophical and political ideal types for a good life (Ec. 2:10; 11:9–10; see also 1:4, 21:9). The first is Socrates, *the philosopher*, who knows "one particular way of making wealth: the generation of surplus" (Xenophon, Ec. 2:10). By moderating his needs, the philosopher can spend most of his time philosophizing (Ec. 2:4). The political ideal type is Socrates' interlocutor, Ischomachos, who is praised as one of "those who are able not only to govern their own oikos but also to accumulate a surplus so that they can adorn the polis and support their friends well; such men must certainly be considered men of strength and abundance" (Xenophon, Ec. 11:9–10). Similarly, these two ideal types of economic behavior are summarized by the lesser-known Arius Didymus (in the 1st century BCE) in the following way: "[A]n oikos [deals] with necessities. These necessities are twofold, those for social life and those for a good life. For the oikonomikos needs first to have forethought about these things, either increasing his revenues through free means of procurement or by cutting down on expenses" (Stobaeus Anthologium II, 7:26).

Thus, the management of the oikos was guided by the ethical disposition that was deemed best-suited to facilitate the engagement of the head of the household in philosophy and politics. Economic theory distinguished between four different

possible ethical dispositions (corresponding to philosophical life, political life, luxurious life, and economic life). It discussed the surplus generated by the economy and the means suited to achieve what was deemed the best ethical disposition.

Abundance, Surplus, and Economic Rationality

The surplus generated by the oikonomia was destined to allow the head of the household to participate in politics and engage in philosophy. If he chose to follow the political ideal type instead of the philosophical one, it also enabled him to be benevolent towards his friends by allowing them leisure time that would enable them to participate in politics and engage in philosophy, as well as supporting the institutions and activities peculiar to the *polis*—that is, the city-state. This perspective is based on three key concepts: abundance, economic rationality, and surplus. *Abundance* is an attribute of nature, which is assumed to be able to meet everyone's needs and beyond—if *economized rationally*. *Surplus*, on the other hand, is the product of people's rational economization of nature's abundance that is not used for securing existence. Thus, the ancient philosophers thought of the oikonomia as a sphere in which man, confronting abundant means, must acquire an ethical disposition of economic rationality enabling him to meet his needs and generate surplus to be spent outside the boundaries of the economic sphere (that is, in philosophy and politics). It is useful to consider these three key closely interrelated components of oikonomia—abundance, economic rationality, and surplus—in more detail.

Abundance

Modern economists hold that means are scarce. However, the ancient Greeks saw nature as potentially capable of satisfying all of man's needs if economized rationally (for an example from Aristotle, see Polanyi 1968, pp. 98–9; in Epicurean and Cynic economics, see Tsouna 2007, pp. 178–80). Moreover, nature was assumed to provide for much more than man's needs, and thus a limit had to be placed on engagement in wealth generation that might otherwise lead men to lose sight of the good life. The need to set a limit to indulgence in wealth generation on the one hand, and the threat of submerging oneself in a luxurious life on the other, meant that nature was seen not just as possessing the means to sustain humanity *abundantly*, but also *excessively*.

Aristotle described how this abundance is economized by the political ideal type saying that "property, in the sense of a bare livelihood, seems to be given by nature herself to all . . . therefore nature makes nothing without purpose or in vain" (Pol. 1256b), and "one kind of art of supply therefore in the order of nature is a part of economics . . . [the] supply of those goods, capable of accumulation, which are necessary for life and useful for the community of *polis* or *oikos*" (Pol. 1256b). Moreover, Aristotle's assertion that nature can supply all of man's needs forms part of his analytic discussion of the science of wealth (referred to as "chrematistics"), in which he discerns between its natural and unnatural kind (for

discussion, see Leshem 2014a). The distinction, according to Aristotle, is that the natural kind is occupied with supplying people's needs, while the unnatural kind (which was presumed to derive from engaging in market trading) is concerned with generating excessive wealth.

Thus, an ancient Greek philosopher generated surplus by restraint in consumption. In her review of Philodemus' *Peri Oikonomia* from the 1st century BCE, Voula Tsouna (2007, p. 182; see also Asmis 2004, p. 145) describes how abundance is economized by the philosopher: "What makes it possible for the philosopher to feel and act in such a way [to be indifferent towards wealth] is, indeed, his confidence that Epicurus was right in saying that natural and necessary desires are easy to satisfy, and their fulfillment is all that the philosopher needs in order to pursue his way of life."

Those who followed the path of virtuous public life sought to generate surplus that could be distributed to the rest of the polis. For example, Bryson the Neo-Pythagorean (163–65) explained: [5]

> [T]he one who conducts himself in this [rational] manner, the fruits and profits of it shall be used for his earnings and enough for the prosperity of his body and food for those in his household, and he should leave some on top of this to help his relatives and acquaintances . . . and a small measure of it to women and the poor people of his *polis* and he should save some so as to be helped in dire times and it is worthy of him not to ask for more than this, and if he asks more than this he is subjected to a bad thing.

Economic Rationality

From the perspective of the ancient Greek philosophers, the problem one is facing when economizing the needs of the oikos is *not* how to deal with the inevitable tradeoffs posed by scarce means. Rather, it is how to set a limit to engaging in economic matters, since nature possess excessive means that can supply all of people's natural needs *as well as* their unnatural desires. On the other hand, if economized rationally, nature can supply the needs of all the inhabitants of one's oikos or polis *and* free some of its members from engaging in economic matters to experience the good life, which is extra-economic.

In this way, what the ancient Greeks meant by rationality in economics clearly differs from the modern view. Modern economics "involves *inter alia* a firm rejection of the 'ethics-related' view" in the words of Amartya Sen (1987, p. 15). In contrast, the ancient Greeks held that the "economy is intelligible only as an ethical domain," which Booth (1993, p. 8) argued is "to be counted among the most significant of their contributions." As a result, the economic approach to human behavior of the ancient Greeks did not begin from an assumption that desires cannot be saturated

[5] The original Greek text (dated to the 1st century CE) was lost. Its translation into Arabic (and later into Hebrew) was preserved. The translation of Bryson is mine, from the Hebrew version.

and therefor scarcity and tradeoffs are inevitable. Instead, they believed that an oikonomia of maximization of desire satisfaction was unethical, and despised those who acted that way. Xenophon (Ec. 1:22) vividly portrayed such people as

> . . . slaves . . . ruled by extremely harsh masters. Some are ruled by gluttony, some by fornication, some by drunkenness, and some by foolish and expensive ambitions which rule cruelly over any men they get into their power, as long as they see that they are in their prime and able to work . . . mistresses such as I have described perpetually attack the bodies and souls and households all the time that they dominate them.

The ancient Greeks saw economic behavior as rational when it was frugal in its use of means towards what they deemed as worthwhile ends. In order to assure the achievement of economic rationality in the sense of the use of means towards praiseworthy ends they appointed the virtue of "soundness of mind" (*sophrosyne*) as the virtue in command of the economy. Aristotle (*Eth. Nic.* 1140b) said that this virtue is called "sophrosyne" because it keeps unharmed (*suzei*) economic rationality (*phronesis*). "Economizing with a sound mind" meant keeping the distinction between needs and desires intact and making sure that the two were incommensurable: needs are to be fully satisfied, while a limit must be set to the otherwise never-ending pursuit of desire gratification. Such an ethical oikonomia generates surplus, and the nature of the surplus generated serves as the ultimate test to the quality of oikonomia.

Moreover, in the literature concerning oikonomia, acquiring a rational disposition was seen as reflecting an ethical choice. This position is very different from contemporary economic theory, which presupposes every economic action as rational without moral qualification and assumes that people's rational disposition can be inferred from their revealed preferences.

Surplus

Perrotta (2004, p. 9) uses the economic concept of surplus, defined as "wealth which exceeds a society's normal consumption," to distinguish between ancient and modern economics. He argues that in modernity the surplus is channeled back into the economic sphere of production, as part of the process of generating economic growth. In contrast, the ancient Greek philosophers distinguished between four uses of surplus (as discussed in Leshem 2013b). The first use of surplus was channeling it back to the economy. This choice was deemed slavish, as it entailed submerging oneself to never-ending economic activity. As such, it missed the end of economic rationality—which was meant to free the head of the household from economic occupations altogether. The latter three uses of surplus are found outside the economic domain and are labeled by Aristotle (Nic. Eth. 1095b) as political, philosophical, and luxurious forms of life. Although a few schools of thought (such as Cynics and Epicureans) disagreed with Aristotle's assertion the good life could only be philosophical or political,

they all agreed that a luxurious life (as well as an unending focus on economic life) is a perversion of the good life.

In most texts, the surplus generated by rational economization of nature's abundance is to be spent beyond the boundaries of the economic sphere of needs satisfaction. At its basic level, the surplus is spent as leisure time (*scholeh*) in which the philosopher surpasses the affairs of this world and ascends into the realm of thought while the citizen participates in politics. The citizen was also expected to use material surplus to demonstrate the virtue of benevolence towards his friends, allowing them leisure time that will enable them to participate in politics and engage in philosophy. Besides sharing the surplus with friends, he was also praised for using it to finance the political institutions.

The Ancient Economics of Human Resources and Property

It was customary to divide the practical discussion of oikonomia, which advised the head of the household how to manage his estate, into four branches: slaves and servants; the wife or matron; children; and property (for examples from various schools, see Leshem 2014b; Natali 1995). Of these four branches, wrote Aristotle: "It is clear then that oikonomia takes more interest in the human members of the oikos than in its inanimate property, and in the excellence of these than in that of its property, which we call riches, and more in that of its free members than in that of slaves" (Pol. 1259b). Or as in the first book of economics: "Of property, the first and most necessary part is that which is best and chiefest; and this is man" (Pseudo Aristotle, Econ. I: 1344a).

The classical oikos was perceived as a partnership between the matron and the master (Xenophon, Ec. 7:12), which "aims not merely at existence, but at a happy existence" (Pseudo Aristotle, Econ. I: 1343b). Of all the actors in ancient economics, the matron demonstrates perhaps the greatest resemblance to contemporary *homo economicus*; unlike the slave, she was a freeborn, and unlike the master, she spent the bulk of her time in the economic domain as she was excluded from the public sphere of politics and was also barred from engaging in philosophy by most schools of thought. As a result, no limit was set on her pursuit of happiness through wealth generation, which took place in the economic sphere. The main difference between the ancient matron and the contemporary *homo economicus* is that the matron was expected to govern the interior of the oikos by demonstrating the virtue of soundness of mind, in which she was seen as either superior to the master (Phyntis: 27) or at least capable of excelling just as much (Xenophon Ec. 7:42). In contrast, the wants of modern contemporary economic man are assumed to know no saturation. As I describe in Leshem (2014b), the matron's pursuit of wealth did not stand in contrast to the master's attempts to set a limit to such an engagement. Rather, the economic harmony between the sexes was conceived as a result of the singular position that the matron occupied in the oikos and the mode by which she demonstrated the virtue of economic soundness of mind. The matron,

as the one entrusted with the economy of preservation, use, and consumption, contributes to wealth generation by efficient inventory management, by rational use of durable goods, and by temperate consumption. Doing all these, she contributed to the generation of extra-economic surplus for her master. As Xenophon's model citizen, Ischomachos, "mansplained" to his young wife: "[M]y bringing in supplies would appear . . . ridiculous if there were not someone to look after what has been brought in. Don't you see how people pity those who draw water in a leaky jar, as the saying goes, because they seem to labor in vain?" (Xenophon Ec. 7:40). Xenophon is pointing out that rational preservation, use, and consumption by the matron can free the master to engage in leisurely occupations such as politics and philosophy.

The ancient Greek philosophers mention children mostly in passing and pay very little attention to economy of the children. When they do discuss the subject, it is usually as part of the husband-wife relationship, as the outcome and purpose of this relationship, and as the ones who will also take care of the parents as they get older.

With the exception of Aristotle, all of the authors dedicate their treatises exclusively to what Aristotle called to the "science of using slaves" (*Politics* 1255b) without expressing any moral qualms about the practice of slavery.[6] The discussion was then divided into three subfields: classification, management, and supervision of slaves. Slaves were classified by Xenophon (Ec. 12–15, 21), Aristotle (Pol. 1255b), and Pseudo Aristotle (Econ. 1: 1344a) into managerial and manual labor. Slaves were managed only for the benefit of their master. Slaves lacked any legal protection, but the ethical disposition of soundness of mind was supposed to stop the master from overusing the slaves and instead cause the master to find a balance between justness and utility.

These texts offer some embryonic discussions of how to set incentives for labor in the context of what we would now call a principal-agent problem. The authors suggest various ways of managing slaves by setting up complex schemes of positive and negative incentives that are meant to make the slaves act in a way that will best serve both their interest and the interest of their master. The incentives recommended were mostly material incentives, and a preference for positive over negative incentives can be easily detected. Theano, for example, justified this preference in her letter to Kallisto on the grounds that "the greatest thing . . . is good will on the slaves' part. For this will is not bought with their bodies." In setting his scheme of incentives for slaves, Xenophon's Ischomachus set negative incentives for conduct he deemed unworthy and positive incentives for conduct he deemed worthy (Ec. 14: 3–6).

[6] It should be noted that Aristotle had no disagreement with or critique of slave labor. However, his purpose in his discussion of slavery was the different one of trying to distinguish between slave by nature and a free man, and between human and nonhuman capital.

Supervision by masters over slaves was also deemed necessary because, as argued by Xenophon, "the master eye produces beautiful and good work" (Ec. 12:20). Pseudo Aristotle (Econ. I: 1345a) describes these in the following manner:

> The master and matron should, therefore, give personal supervision, each to his or her special department of the oikonomia. In small oikonomiai, an occasional inspection will suffice; in ones managed through stewards, inspections must be frequent. For in stewardship as in other matters there can be no good copy without a good example; and if the master and matron do not attend diligently [to their oikonomia], their deputies will certainly not do so.

As a result of the emphasis of the ancient Greeks on human resources, the economy of property is barely discussed. Most of their discussion aims over and again at defining the proper limit to the production and accumulation of wealth, either for the political or the philosophical ideal type. The discussion of methods of production, distribution, and accumulation, once the proper limit has been set, is rather dull. In general, it does not go beyond prosaic advice such as "the oikonomos must . . . have the faculty of acquiring, and . . . that of preserving what he has acquired; otherwise there is no more benefit in acquiring than in baling with a colander, or in the proverbial wine-jar with a hole in the bottom" (Pseudo-Aristotle, Econ. I: 1344b). Philodemus of Gadara, the only author who dedicates his book solely to property oikonomia, essentially focuses on offering a critique of the commonly held view that one should maintain a fixed level of expenditure and spread one's investment in order to minimize risk. Instead, he argues for more flexibility in asset management on the philosopher's behalf (Philodemus 2012: 30–32).

Oikonomia, Ethics, and Modern Economics

Of course, pointing out that ancient oikonomia is intertwined with ethical judgments gives no assurance that ancient economics stands on higher moral grounds than the modern study of economics, which largely dissociated itself from ethical considerations. The surplus of ancient "ethical" oikonomia was generated by slave labor and the denial of citizen rights to women. It was the abuse of slaves alongside the continuous subjection of female citizens that enabled "all people rich enough to be able to avoid personal trouble [to] have a steward who takes this office, while they themselves engage in politics or philosophy" (Aristotle, Pol. 1255b). At the same time, the culpability of the ancient oikonomia need not imply the desirability of an approach to economics that is built on a separation from ethics. Surely, we can think of ways to engage in ethical economics without slavery or the denying of human and civil rights.

One recent attempt to rejoin economics and ethics is Amartya Sen's "capability approach." As Sen (1993) notes, his approach has links to Aristotle's understanding of human flourishing. Sen's approach argues for assessing the performance of the

economy based on people's "capability" to attend to "functionings." The former includes both life necessities such as access to food and shelter, as well as access to functionings necessary for what the ancient Greek philosophers deemed as prerequisites for a good life, such as access to literacy and participation in democracy. The functionings sought after are not solely based on people's subjective assessments of their own situation as with approaches based on ordinal utility or, more recently, happiness indices. This is because people who face lives of deprivation, sickness, and limited opportunities may not be able to know or to enunciate what they are capable of, or what they should want. Sen's approach is also different from indices that measure the overall performance of the economy in terms of aggregate GDP. Sen's approach is not indifferent to how income is distributed among the members of society or the extent to which people have basic human and civil rights. Much like the ancient Greek philosophers, Sen's vision of capabilities is not neutral between ends. Sen abstains from enunciating a precise and explicit definition of what functionings should count as necessary for a good life, in part because he is taking into account the extent to which perceptions of this may vary across countries with different income levels and cultural traditions.

Of course, one can also suggest a variety of other ethical underpinnings for a modern economics. But many of these approaches would argue that the ends of economic analysis should be open to an ethical discussion and that economic rationality should be defined in terms of how best to approach the goals that emerge from an ethical framework. Indeed, as many parts of the world attain an ever-higher state of economic progress, an ethical framework might call into question the pursuit of economic goals as an end in and for themselves. At least in this sense, the ancient ethical oikonomia—stripped of the abusive qualities characteristic of its time—may serve as a source of inspiration for seeking to mix the practicalities of economic life with an articulated ethics of human purpose.

References

Ancient Sources

Aesara. 1987. "On Human Nature." In *A History of Women Philosophers*, Vol. 1: *Ancient Women Philosophers, 600 B.C.–500 A.D.*, edited by Mary Ellen Waithe, 20–21. Dordrecht: Kluwer Academic.

Aristotle. 1934. *The Nicomachean Ethics*. Volume 19 of *Aristotle in 23 Volumes*. Translated by H. Rackham. Cambridge, MA: Harvard University Press.

Aristotle. 1944. *Politics*. Volume 21 of *Aristotle in 23 Volumes*. Cambridge MA: Harvard University Press. Translated by H. Rackham.

Bryson. 1928. *Der Oikonomikos des Neupythagoreers Bryson und sein Einfluss auf die islamische Wissenschaft*, edited by Martin Plessner. Heidelberg: C. Winter.

Callicratidas. 1987. "On the Felicity of Families." In *The Pythagorean Sourcebook and Library: An Anthology of Ancient Writings which relate to Pythagoras and Pythagorean Philosophy*, compiled

and translated by Kenneth Slyvan Guthrie, 235–237. Grand Rapids, MI: Phanes Press.

Didymus, Arius. 1999. *Epitome of Stoic Ethics*, edited by Arthur J. Pomeroy. Atlanta, GA: Society of Biblical Literature.

Hesiod. 1914. "Works and Days." In *The Homeric Hymns and Homerica with an English Translation by Hugh G. Evelyn-White*. Cambridge, MA. Harvard University Press; London, William Heinemann Ltd.

Hierocles. 1987. "On Economics." In *The Pythagorean Sourcebook and Library: An Anthology of Ancient Writings which relate to Pythagoras and Pythagorean Philosophy,* edited by Kenneth Sylvan Guthrie, 285–86. Grand Rapids, MI: Phanes Press.

Laertius, Diogenes. 1853. *The Lives and Opinions of Eminent Philosophers,* translated by C .D Younge. London: George Bell & Sons.

Lysias. 1930. "On the Murder of Eratosthenes." In *Lysias. With an English Translation by W. R. M. Lamb.* Cambridge, MA: Harvard University Press.

Myia. 1987. "Letter to Phyllis." In *A History of Women Philosophers*, Vol. 1: *Ancient Women Philosophers, 600 B.C.–500 A.D.*, edited by Mary Ellen Waithe, 15–17. Dordrecht: Kluwer Academic.

Perictione-I. 1987. "On the Harmony of Woman." In *Ancient Women Philosophers, 600 B.C.–500 A.D.*, edited by Mary Ellen Waithe, 32–34. Dordrecht: Kluwer Academic.

Philodemus. 2012. *On Property Management.* Translated with introduction and notes by Voula Tsoua. Atlanta: Society of Biblical Literature.

Phocilides. Undated. "In Women in Classical Antiquity: Ancient and Modern Sources, Ancient and Modern Attitudes (Selections)," edited by Judith P. Hallett. The Stoa Consortium, http://www.stoa.org/diotima/anthology/attitudes.shtml.

Phyntis. 1987. "On the Moderation of Women, Fragment 1." In *A History of Women Philosophers*, Vol. 1: *Ancient Women Philosophers, 600 B.C.–500 A.D.*, edited by Mary Ellen Waithe, 26–28. Dordrecht: Kluwer Academic.

Plato. 1966. "Apology." Volume 1 in *Plato in Twelve Volumes*. Translated by Harold North Fowler; introduction by W. R. M. Lamb. Cambridge, MA; Harvard University Press.

Pseudo-Aristotle. 1910. "Economics Book I." *The Politics and Economics of Aristotle*, edited by Edward Walford, 289–303. London: G. Bell & Sons.

Sophocles. 1894. *The Electra of Sophocles.* Edited with introduction and notes by Sir Richard Jebb. Cambridge: Cambridge University Press.

Stobaeus, Ioannis. 1884–1912. *Anthologium*, edited by Curtius Wachsmuth and Otto Hense. Berlin: Weidmann

Theano-II. 1987a. "To Kallisto." In *A History of Women Philosophers*, Vol. 1: *Ancient Women Philosophers, 600 B.C.–500 A.D.*, edited by Mary Ellen Waithe, 47–48. Dordrecht: Kluwer Academic.

Theano-II. 1987b. "To Euboule." In *A History of Women Philosophers*, Vol. 1: *Ancient Women Philosophers, 600 B.C.–500 A.D.*, edited by Mary Ellen Waithe, 42–43. Dordrecht: Kluwer Academic.

Theano-II. 1987. "To Nikostrate." In *A History of Women Philosophers*, Vol 1: *Ancient Women Philosophers, 600 B.C.–500 A.D.*, edited by Mary Ellen Waithe, 44–46. Dordrecht: Kluwer Academic.

Xenophon. 1994. *Oeconomicus: A Social and Historical Commentary.* With a New English translation by Sarah B. Pomeroy. Oxford: Clarendon Press.

Modern Sources

Asmis, Elizabeth. 2004. "Epicurean Economics." In *Philodemus and the New Testament World*, edited by John T. Fitzgerald, Dirk Obbink, and Glenn S. Holland, 133–76. Leiden: Brill.

Baloglou, Christos P. 2009. "The Reception of the Ancient Greek Economic Ideas by the Romans and Their Contribution to the Evolution of Economic Thought." In *Theorie und Geschichte der Wirtschaft: Festschrift für Bertram Schefold,* edited by Caspari, Volker, 191–256. Marburg: Metropolis-Verlang.

Booth, William James. 1993. *Households: The Moral Architecture of the Economy.* Ithaca, NY: Cornell University Press.

Friedman, Milton. 1953. "The Methodology of Positive Economics." *Essays in Positive Economics,* 3–34. Chicago: Chicago University Press.

Leshem, Dotan. 2013a. "Aristotle Economizes the Market." *boundary-2* 40(3): 39–57.

Leshem, Dotan. 2013b. "Oikonomia in the Age of Empires." *History of the Human Sciences* 26(1): 29–51.

Leshem, Dotan. 2013c. "Oikonomia Redefined." *Journal of the History of Economic Thought* 35(1): 43–61

Leshem, Dotan. 2014a. "The Distinction between the Economy and Politics in Aristotle's Thought and the Rise of the Social." *Constellations*, online first, December 4. http://onlinelibrary.wiley.com/doi/10.1111/1467-8675.12128/abstract.

Leshem, Dotan. 2014b. "The Ancient Art of Economics." *European Journal for the History of Economic Thought* 21(2): 201–229.

Natali, Carlo. 1995. "*Oikonomia* in Hellenistic Political Thought." In *Justice and Generosity: Studies in Hellenistic Social and Political Philosophy: Proceedings of the Sixth Symposium Hellenisticum,* edited by Andre Laks and Malcolm Schofield, 95–128. Cambridge: Cambridge University Press.

Perrotta, Cosimo. 2004. *Consumption as an*

Investment: I. The Fear of Goods from Hesiod to Adam Smith. London and New York: Routledge.

Polanyi, Karl. 1968. "Aristotle Discovers the Economy: Essays of Karl Polanyi." In *Primitive, Archaic, and Modern Economics*, edited by George Dalton, 78–115. Garden City: Doubleday.

Reumann, John Henry. 1957. *The Use of Oikonomia' and Related Terms in Greek Sources to about 100 A.D. as a Background for Patristic Applications*. Dissertation, January 1, University of Pennsylvania. http://repository.upenn.edu/dissertations/AAI0023631.

Robbins, Lionel. 1935. *Essay on the Nature and Significance of Economic Science*. 3rd ed. London: The Macmillan Press, Ltd.

Sen, Amartya. 1987. *On Ethics and Economics*. Oxford: Basil Blackwell.

Sen, Amartya K. 1993. "Capability and Well-being." In *The Quality of Life*, edited by Martha C. Nussbaum and Amartya K. Sen, pp. 30–53. Oxford: Clarendon Press.

Trever, Albert Augustus. 1916. *A History of Greek Economic Thought*. Chicago: The University of Chicago Press.

Tsouna, Voula. 2007. *The Ethics of Philodemus*. Cambridge: Cambridge University Press.

Waithe, Mary Ellen. 1987. *Ancient Women Philosophers, 600 B.C.–500 A.D.* Dordrecht: Kluwer Academic.

Journal of Economic Perspectives—Volume 30, Number 1—Winter 2016—Pages 239–246

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by email at taylort@macalester.edu, or c/o *Journal of Economic Perspectives*, Macalester College, 1600 Grand Ave., Saint Paul, Minnesota, 55105.

Speeches

Alan B. Krueger delivered the 2015 Martin Feldstein Lecture at the National Bureau of Economic Research on the subject "How Tight Is the Labor Market?" "By 2013 short-term unemployment had returned to normal levels. So at that time I argued that if we were going to make further progress in lowering the unemployment rate, it would be because the long-term unemployed either found jobs or left the labor force. … Unfortunately, as we will see, the historical pattern in which the long-term unemployed tend to increase their labor force exit rate over the course of the business cycle has reasserted itself during the current recovery … Today, the labor force participation rate is nearly 5 percentage points below its peak. Sensible

■ *Timothy Taylor is Managing Editor,* Journal of Economic Perspectives*, based at Macalester College, Saint Paul, Minnesota. He blogs at http://conversableeconomist.blogspot.com.*

http://dx.doi.org/10.1257/jep.30.1.239 doi=10.1257/jep.30.1.239

analyses suggest that about half of the 15-year decline in labor force participation is due to predictable demographic changes, particularly the aging of the Baby Boom generation. As for the other half, I think there are two important factors. About half of this remainder (or a quarter of the overall decline) can be accounted for by trends that were taking place before the Great Recession and likely continued after it. For instance, the widespread entrance of women into the workforce that had fueled the great postwar rise in labor force participation in the United States peaked around 2000. … In addition, labor force participation of younger workers declined in conjunction with an increase in their school enrollment, which should be a net positive for the economy in the long run. The remaining quarter—or a little over a percentage point—of the overall decline in labor force participation is likely attributable to cyclical factors. I will present evidence suggesting that it's unlikely we'll see much of a recovery for this segment of the population going forward." An edited version of this talk is published in the NBER Digest (2015, number 3, pp. 1–10) at http://www.nber.org/reporter/2015number3/2015number3.pdf. Alternatively, you can watch the talk and access the slides at http://www.nber.org/feldstein_lecture_2015/feldsteinlecture_2015.html.

Ben Bernanke gave the Mundell–Fleming Lecture at the IMF's 16th Jacques Polak Annual Research Conference on November 5, 2015, on the subject, "Federal Reserve Policy in an International Context." "I don't think that US trading partners have much basis, either theoretical or empirical, to complain about currency wars being waged by the Fed. US growth during the recent recovery has certainly not been driven by exports, and, as I will explain, the "expenditure-augmenting" effects of US monetary policies (adding to global aggregate demand) tend to offset the "expenditure-switching" effects (adding to demand in one country at the expense of others). … Regarding financial stability spillovers: Research has documented the strong co-movement of asset prices, credit growth, and leverage across economies, but only limited progress has been made in determining the degree that this co-movement is in some sense excessive or in documenting the channels through which the putative spillovers operate. … I argue that monetary and exchange-rate policies should focus on macroeconomic objectives, with the problem of spillovers being tackled by regulatory and macroprudential measures, possibly including targeted capital controls, and through careful sequencing of market reforms. … My short review of the dollar standard suggests that its benefits to the United States and to US trading partners are much better balanced than in the Bretton Woods era, the days of America's "exorbitant privilege." … The phrase was coined in 1965 by French finance minister Valéry Giscard d'Estaing to describe the gains to the United States from the central role of the dollar in the Bretton Woods system. … [W]e shouldn't be overly exercised over controversies about whether the dollar will retain its pre-eminence, the future of the renminbi as a reserve currency, and so on. These debates are more about symbolism than substance. In purely economic terms, the universal usage of English, say, is far more valuable to the United States than the broad use of the dollar." http://www.imf.org/external/np/res/seminars/2015/arc/pdf/Bernanke.pdf. Video of the

lecture is available at http://www.imf.org/external/mmedia/view.aspx?vid=4598935280001.

Rebecca M. Blank discusses "What Drives American Competitiveness?" in her 2015 Daniel Patrick Moynihan Lecture on Social Science and Public Policy. "The U.S. economy is growing more slowly, and the growth that we have experienced is not translating into higher incomes. These problems emerged in the early 2000s, before the Great Recession, and seem to be continuing even as the U.S. economy is close to fully recovered from that economic downturn. Furthermore, the slowdown in productivity suggests that our prospects for future growth are limited … Between 1970 and 1999, growth in hours worked was a key factor for total growth, accounting for around 1 percentage point of GDP growth each year. This factor has fallen sharply since 2000, and has actually pulled down growth a small amount in the past 15 years. Growth in labor skills is the smallest factor in GDP growth over this time period. Labor skills showed little growth in the 1970s but have accounted for between 0.3 and 0.4 percentage points of GDP growth in the decades since. The two innovation factors have been important throughout the past 45 years. Capital deepening is the most important factor in every time period, accounting for more than 1 percentage point growth in GDP from 1970 to 1999 and only a little less in the last 15 years. Total Factor Productivity is slightly less important than capital depending in most decades." *Annals of the American Academy of Political and Social Science*, January 2016, 663, pp. 8–30. A link to view the lecture is available at http://www.aapss.org/the-moynihan-prize.

Volumes

Richard Baldwin and Francesco Giavazzi have edited *The Eurozone Crisis: A Consensus View of the Causes and a Few Possible Solutions*, which includes their short introduction and 14 short and all readable essays. Here's the capsule summary from the introduction: "The core reality behind virtual every crisis is the rapid unwinding of economic imbalances. ... From the euro's launch and up until the crisis, there were big capital flows from EZ [eurozone] core nations like Germany, France, and the Netherland to EZ periphery nations like Ireland, Portugal, Spain and Greece. A major slice of these were invested in non-traded sectors—housing and government services/consumption. This meant assets were not being created to help pay off in the investment. It also tended to drive up wages and costs in a way that harmed the competitiveness of the receivers' export earnings, thus encouraging further worsening of their current accounts. When the EZ crisis began—triggered ultimately by the Global Crisis—cross-border capital inflows stopped. This 'sudden stop' in investment financing raised concerns about the viability of banks and, in the case of Greece, even governments themselves. The close links between EZ banks and national governments provided the multiplier that made the crisis systemic. Importantly, the EZ crisis should not be thought of as a sovereign debt crisis. … The key was foreign borrowing. Many of the nations that ran current account deficits—and thus

were relying of foreign lending—suffered; none of those running current account surpluses were hit." 2015. A VoxEU.org book from the Centre for Economic Policy Research in London. http://www.voxeu.org/sites/default/files/file/reboot_upload_0.pdf.

The Future of Children has devoted an issue with eight essays to the theme "Marriage and Child Wellbeing Revisited." From the overview by Sara McLanahan and Isabel Sawhill: "Marriage is on the decline. Men and women of the youngest generation are either marrying in their late twenties or not marrying at all. Childbearing has also been postponed, but not as much as marriage. The result is that a growing proportion of children are born to unmarried parents—roughly 40 percent in recent years, and over 50 percent for children born to women under 30. Many unmarried parents are cohabiting when their child is born. Indeed, almost all of the increase in nonmarital childbearing during the past two decades has occurred to cohabiting rather than single mothers. But cohabiting unions are very unstable, leading us to use the term "fragile families" to describe them. About half of couples who are cohabiting at their child's birth will split by the time the child is five. Many of these young parents will go on to form new relationships and to have additional children with new partners. The consequences of this instability for children are not good. Research increasingly shows that family instability undermines parents' investments in their children, affecting the children's cognitive and social-emotional development in ways that constrain their life chances." Fall 2015, http://www.futureofchildren.org/futureofchildren/publications/journals.

Smorgasbord

The Committee for the Prize in Economic Sciences in Memory of Alfred Nobel explained the choice of its 2015 award in the essay "Angus Deaton: Consumption, Poverty and Welfare." "Over the last three to four decades, the study of consumption has progressed enormously. While many scholars have contributed to this progress, Angus Deaton stands out. … His main achievements are three. First, Deaton's research brought the estimation of demand systems—i.e., the quantitative study of consumption choices across different commodities—to a new level of sophistication and generality. The Almost Ideal Demand System that Deaton and John Muellbauer introduced 35 years ago, and its subsequent extensions, remain in wide use today—in academia as well as in practical policy evaluation. Second, Deaton's research on aggregate consumption helped break ground for the microeconometric revolution in the study of consumption and saving over time. He pioneered the analysis of individual dynamic consumption behavior under idiosyncratic uncertainty and liquidity constraints. He devised methods for designing panels from repeated cross-section data, which made it possible to study individual behavior over time, in the absence of true panel data. … Third, Deaton spearheaded the use of household survey data in developing countries, especially data on consumption, to measure living standards and poverty. In so doing, Deaton helped transform development

economics from a largely theoretical field based on crude macro data, to a field dominated by empirical research based on high-quality micro data." 2015. http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2015/advanced-economicsciences2015.pdf. For a brief and readable overview of this longer paper, see the Committee's "Information for the Public: Consumption, great and small" at http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2015/popular-economicsciences2015.pdf.

The World Trade Organization devotes its *World Trade Report 2015* to the theme, "Speeding Up Trade: Benefits and Challenges of Implementing the WTO Trade Facilitation Agreement." "While trade agreements in the past were about "negative" integration—countries lowering tariff and non-tariff barriers—the WTO Trade Facilitation Agreement (TFA) is about positive integration—countries working together to simplify processes, share information, and cooperate on regulatory and policy goals. … The TFA represents a landmark achievement for the WTO, with the potential to increase world trade by up to US$ 1 trillion per annum. … Based on the available evidence, trade costs remain high. … [T]rade costs in developing countries in 2010 were equivalent to applying a 219 per cent ad valorem tariff on international trade. This implies that for each dollar it costs to manufacture a product, another US$ 2.19 will be added in the form of trade costs. Even in high-income countries, trade costs are high, as the same product would face an additional US$ 1.34 in cost." 2015, https://www.wto.org/english/res_e/publications_e/wtr15_e.htm.

A High Level Panel on Humanitarian Cash Transfers, which includes a mixture of academics and those working in development, have produced the report "Doing Cash Differently: How Cash Transfers Can Transform Humanitarian Aid." "Give more unconditional cash transfers. The questions should always be asked: 'why not cash?' and, 'if not now, when?'." "[T]he Panel estimates that cash and vouchers together have risen from less than 1% in 2004 to around 6% of total humanitarian spending today … If sectors where cash is often less appropriate (health, water and sanitation) and not appropriate at all (mine action, coordination, security) are removed from the equation, then cash and vouchers were roughly 10% of the total." "A consistent theme in research and evaluations is the flexibility of cash transfers, enabling assistance to meet a more diverse array of needs. In the Philippines, for example, people reported using the money for food, building materials, agricultural inputs, health fees, school fees, sharing, debt repayment, clothing, hygiene, fishing equipment and transport. … The element of choice is critical. … Cash impacts local economies and market recovery by increasing demand and generating positive multiplier effects. In Zimbabwe, every dollar of cash transfers generated $2.59 in income (compared to $1.67 for food aid). It can encourage the recovery of credit markets by enabling repayment of loans." "It usually costs less to get money to people than in-kind assistance because aid agencies do not need to transport and store relief goods. A four-country study comparing cash transfers and food aid found that 18% more people could be assisted at no extra cost if everyone received cash instead of food." "Providing cash does not and should not mean that humanitarian actors lose a focus on a

key public good that they are uniquely placed to provide: proximity, presence and bearing witness to the suffering of disaster-affected populations. On the contrary, streamlining aid delivery should allow them more time to focus on exactly that. Giving people cash, therefore, does not imply simply dumping the money and leaving them to fend for themselves. People receiving cash intended to help meet shelter needs may require help to secure land rights, build disaster-resistant housing or manage procurement and contractors. Where people use cash to buy agricultural inputs this can be complemented with extension advice." September 2015, published by the Center for Global Development, http://www.cgdev.org/sites/default/files/HLP-Humanitarian-Cash-Transfers-Report.pdf.

The Council of Economic Advisers summarizes research on unions and wages in an Issue Brief titled "Worker Voice in a Time of Rising Inequality." "The overall effect of unions on the wage distribution depends on both the union wage premium and the types of workers who are unionized. Unionized workers still command a sizable wage premium of up to 25 percent relative to similar nonunionized workers, but that premium has fallen slightly over the past couple of decades. Union membership has also become more representative of the population, with the share of members who are female or college-educated rising quickly. Studies have shown that union wage effects are largest for workers with low levels of observed skills and that unionization can reduce wage inequality among workers partially by increasing wages at the bottom of the distribution and by reducing pay dispersion within unionized firms and industries. Since both the union wage premium and the coverage of low-skilled workers, who receive the highest wage premium, have fallen, unionization's ability to reduce inequality has very likely been limited in recent years." October 2015, https://www.whitehouse.gov/sites/default/files/page/files/worker_voice_issue_brief_cea.pdf.

Irwin Garfinkel and Timothy Smeeding challenge three claims in their essay, "Welfare State Myths and Measurement": 1) "The current American welfare state is unusually small." 2) "The United States has always been a welfare state laggard." 3) "The welfare state undermines productivity and economic growth." They write: "Very reasonable changes in measurement reveal that all three beliefs are untrue." For example, they look at the US welfare state in absolute per capita terms and write: "The United States, as one of the richest nations, could be spending more in absolute terms and less as a percentage of income than other rich nations. … Australia, for example, spent a slightly larger proportion of its GDP on SWE [social welfare expenditures] in 2001 than the US ... but its [per capita] GDP then was only a bit above 60% of US GDP. Consequently, US per capita social welfare expenditures are much higher than Australia's. … Real per capita social welfare spending in the United States is larger than that in almost all other countries! Even if employer-provided benefits and tax expenditures are excluded, the United States is still the third biggest spender on a per capita basis." *Capitalism and Society* (vol. 10, no. 1), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2629585.

David Price has an interview with James Poterba. On changes in tax policy research: "One difference is that tax policy discussions and research on the economics of tax

policy in the late 1970s and early 1980s were set in an environment with marginal tax rates that were significantly higher than those today. The United States had a top tax rate on capital income of 70 percent until 1981. The top marginal tax rate on earned income in the United States at the federal level was 50 percent until 1986. Today, the top statutory rate is 39.6 percent, although with some add-on taxes, the actual rate can be in the low 40s. We have been through periods when the top rate was as low as 28 percent. There was a lot more concern about the distortions associated with the capital income tax and with taxation in general. … We relied primarily on cross-sectional household surveys. It's hard to study how taxation affects behavior when the variation in the tax system is coming in differences in household incomes that place different taxpayers in different tax brackets, because income variation is related to so many other characteristics. Today, by comparison, the field of public finance has moved forward to use large administrative databases from many countries, often including tax returns. It is possible to do a much more refined kind of empirical analysis than when I started." On retirement saving: "The University of Michigan Health and Retirement Study, which is a comprehensive database on older individuals in the United States, begins tracking survey respondents in their mid-50s. It follows them until they die, so the last survey is typically filed about a year before the individual's death. Nearly half of the respondents in the survey turn out to have very low levels of financial assets, under $20,000, as they get close to death. … I have been quite interested in how individuals arrive at such low levels of financial assets. Many of those who have very little financial wealth as they approach death also reached retirement age with very little wealth. … Only in the top half of the retiree wealth distribution does one start to see substantial amounts of support from private pension plans, and only in the top quarter is there substantial support from private saving outside retirement accounts." *Econ Focus,* Federal Reserve Bank of Richmond, 2015, Second Quarter, pp. 24–29, https://www.richmondfed.org/publications/research/econ_focus/2015/q2/interview.

Discussion Starters

P. J. Held, F. McCormick, A. Ojo and J. P. Roberts present "A Cost–Benefit Analysis of Government Compensation of Kidney Donors." "From 5000 to 10 000 kidney patients die prematurely in the United States each year, and about 100 000 more suffer the debilitating effects of dialysis, because of a shortage of transplant kidneys. To reduce this shortage, many advocate having the government compensate kidney donors. This paper presents a comprehensive cost–benefit analysis of such a change. It considers not only the substantial savings to society because kidney recipients would no longer need expensive dialysis treatments—$1.45 million per kidney recipient—but also estimates the monetary value of the longer and healthier lives that kidney recipients enjoy—about $1.3 million per recipient. These numbers dwarf the proposed $45 000-per-kidney compensation that might be needed to end the kidney shortage and eliminate the kidney transplant waiting

list. From the viewpoint of society, the net benefit from saving thousands of lives each year and reducing the suffering of 100 000 more receiving dialysis would be about $46 billion per year, with the benefits exceeding the costs by a factor of 3. In addition, it would save taxpayers about $12 billion each year." *American Journal of Transplantation*, published online October 16, 2015, at http://onlinelibrary.wiley.com/doi/10.1111/ajt.13490/full. The article can be a useful follow-up to the essay by Gary S. Becker, and Julio Jorge Elías, "Introducing Incentives in the Market for Live and Cadaveric Organ Donations," which appeared in the Summer 2007 issue (pp. 3–24) of this journal.

The McKinsey Global Institute reports on *The Power of Parity: How Advancing Women's Equality Can Add $12 Trillion to Global Growth."* "We also analyzed an alternative "best-in-region" scenario in which all countries match the rate of improvement of the best-performing country in their region. This would add as much as $12 trillion in annual 2025 GDP, equivalent in size to the current GDP of Japan, Germany, and the United Kingdom combined, or twice the likely growth in global GDP contributed by female workers between 2014 and 2025 in a business-as-usual scenario … with the highest relative boost in India and Latin America. … Seventy-five percent of the world's total unpaid care is undertaken by women, including the vital tasks that keep households functioning such as child care, caring for the elderly, cooking, and cleaning. However, this contribution is not counted in traditional measures of GDP. Using conservative assumptions, we estimate that unpaid work being undertaken by women today amounts to as much as $10 trillion of output per year, roughly equivalent to 13 percent of global GDP." September 2015. http://www.mckinsey.com/insights/growth/how_advancing_womens_equality_can_add_12_trillion_to_global_growth.

Journal of Economic Perspectives—Volume 30, Number 1—Winter 2016—Pages 247–248

Correspondence

To be considered for publication in the Correspondence section, letters should be relatively short—generally less than 1,000 words—and should be sent to the journal offices at jep@jepjournal.org. The editors will choose which letters will be published. All published letters will be subject to editing for style and length.

The Doing Business Project: How It Started

The Summer 2015 issue included an article by Timothy Besley on the nature and influence of the World Bank's Doing Business report ("Law, Regulation, and the Business Climate: The Nature and Influence of the World Bank Doing Business Project," pp. 99–120). As the manager of the World Bank team that created Doing Business, I wish to highlight the importance of academic research in starting this project.

The Doing Business report was first published in 2003 with five sets of indicators for 133 economies. However, the team that created Doing Business had been formed three years earlier, during the writing of the *World Development Report 2002: Building Institutions for Markets* (World Bank 2001). The focus on the importance of institutions in development was chosen by Joseph Stiglitz, who at the time was the World Bank's Chief Economist. As a member of the team, I was tasked with authoring the chapters on institutions and firms. At the time, the work by Rafael La Porta, Florencio Lopez de Silanes, Andrei Shleifer, and Robert Vishny on legal origins and various aspects of institutional evolution was generating a great deal of interest. I turned to Shleifer with a request to collaborate on several background papers for the World Development Report. He agreed, on the condition that we used this work as an opportunity to gather and analyze new cross-country datasets on institutions. This is how Doing Business started.

The inspiration behind Doing Business was two-fold. First, both Shleifer and I had previously researched the experience of centrally planned economies and documented the waste of entrepreneurial talent and resources as a result of overregulation. With the collapse of communism, research on the benefits of simpler regulation would be of use to reformers in Eastern Europe. Second, in his book *The Other Path*, Hernando de Soto (1989) showed that the prohibitively high cost of establishing a business in Peru denies economic opportunity to the poor.

Working with Shleifer, we constructed five sets of indicators on the institutions that affect entrepreneurs and businesses through their life-cycle, the results of which were published in a series of papers in academic journals: 1) Djankov, La Porta, Lopes-de-Silanes, and Shleifer (2002) on starting a business, with measures of the procedures, time, cost, and minimum capital required to start a new business; 2) Djankov, La Porta, Lopez-de-Silanes, and Shleifer (2003) on enforcing contracts, with measures of the procedures, time, and cost required to enforce a debt contract; 3) Djankov, McLiesh, and Shleifer (2007) on getting credit, with measures of the strength of legal rights, which encompass the degree to which collateral and bankruptcy laws protect the rights of borrowers and lenders, and the depth of sharing credit information; 4) Botero, Djankov, La Porta, Lopez-De-Silanes, and Shleifer (2004) on employing workers, with measures of the ease with which workers can be hired or made redundant and the rigidity of working hours; and 5) Djankov, Hart, McLiesh, and Shleifer (2008) on resolving insolvency, with measures of the time, cost, and percentage recovery rate involved with bankruptcy proceedings.

The data collection started simultaneously on all five projects and was based on a reading of the laws and regulations. The measures of institutional efficiency and quality were intended to be comparable across countries, which was achieved by basing the

data collection on a precisely defined hypothetical enterprise and the circumstances that it faced. The hypothetical case is a firm with at least 60 employees, which is located in the country's largest business city. It is a private, limited-liability company and does not operate in an export-processing zone or an industrial estate with special export or import privileges. It is 100 percent domestically owned, and exports constitute more than 10 percent of its sales.

The five papers were extensively used in working paper form in the World Development Report and were then picked up as the basis for a new ongoing effort which became the Doing Business project. World Bank Vice-President Michael Klein provided the needed guidance for the project to develop and survive various challenges within the World Bank. The set of Doing Business indicators gradually expanded from five to eight, using essentially the same methodology. The new additions were published as: 6) Djankov, Freund, and Pham (2010) on additional sets of measures on the efficiency of customs rules; 7) Djankov, Ganser, McLiesh, Ramalho, and Shleifer (2010) on the effect of different tax regimes on entrepreneurship; and 8) Djankov, La Porta, Lopez-de-Silanes, and Shleifer (2008) on investor protections.

The high visibility and impact of the Doing Business project generated controversy reflected in discussions at the World Bank and by its shareholders. This resulted in one set of indicators—on employing workers—being removed from the overall Doing Business ranking. Another proposed set of indicators on disclosure of assets and income by politicians was not taken up by the World Bank (Djankov, La Porta, Lopez-de-Silanes, and Shleifer 2010).

Simeon Djankov
Finance Department, London School of Economics
London, United Kingdom
S.Djankov@lse.ac.uk

References

Botero, Juan C., Simeon Djankov, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2004. "The Regulation of Labor." *Quarterly Journal of Economics* 119(4): 1339–82.

de Soto, Hernando. 1989. *The Other Path.* New York: Harper and Row.

Djankov, Simeon, Caroline Freund, and Cong S. Pham. 2010. "Trading on Time," *Review of Economics and Statistics* 92(1): 166–73.

Djankov, Simeon, Tim Ganser, Caralee McLiesh, Rita Ramalho, and Andrei Shleifer. 2010. "The Effect of Corporate Taxes on Investment and Entrepreneurship." *American Economic Journal: Macroeconomics* 2(3): 31–64.

Djankov, Simeon, Oliver Hart, Caralee McLiesh, and Andrei Shleifer. 2008. "Debt Enforcement around the World." *Journal of Political Economy* 116(6): 1105–49.

Djankov, Simeon, Rafael La Porta, Florencio Lopes-de-Silanes, and Andrei Shleifer. 2002. "The Regulation of Entry." *Quarterly Journal of Economics* 117(1): 1–37.

Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2003. "Courts." *Quarterly Journal of Economics* 118(2): 453–517.

Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2008. "The Law and Economics of Self-Dealing." *Journal of Financial Economics* 88(3): 430–65.

Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2010. "Disclosure by Politicians." *American Economic Journal: Applied Economics* 2(2): 179–209.

Djankov, Simeon, Caralee McLiesh, and Andrei Shleifer. 2007. "Private Credit in 129 Countries." *Journal of Financial Economics* 84(2): 299–329.

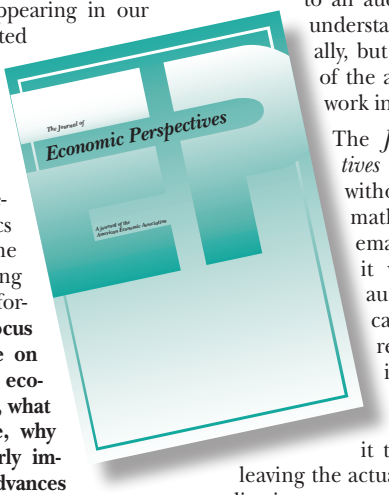
World Bank. 2001. *World Development Report 2002: Building Institutions for Markets.* Washington DC.

AMERICAN
ECONOMIC
ASSOCIATION

The *Journal of Economic Perspectives*: Proposal Guidelines

Considerations for Those Proposing Topics and Papers for *JEP*

Articles appearing in the journal are primarily solicited by the editors and associate editors. However, we do look at all unsolicited material. Due to the volume of submissions received, proposals that do not meet *JEP*'s editorial criteria will receive only a brief reply. Proposals that appear to have *JEP* potential receive more detailed feedback. Historically, about 10–15 percent of the articles appearing in our pages originate as unsolicited proposals.



Philosophy and Style

The *Journal of Economic Perspectives* attempts to fill part of the gap between refereed economics research journals and the popular press, while falling considerably closer to the former than the latter. **The focus of *JEP* articles should be on understanding the central economic ideas of a question, what is fundamentally at issue, why the question is particularly important, what the latest advances are, and what facets remain to be examined.** In every case, articles should argue for the author's point of view, explain how recent theoretical or empirical work has affected that view, and lay out the points of departure from other views.

We hope that most *JEP* articles will offer a kind of intellectual arbitrage that will be useful for every economist. For many, the articles will present insights and issues from a specialty outside the readers' usual field of work. For specialists, the articles will lead to thoughts about the questions underlying their research, which directions have been most productive, and what the key questions are.

Articles in many other economics journals are addressed to the author's peers in a subspecialty; thus, they use tools and terminology of that specialty and presume that readers know the context and general direction of the inquiry. By contrast, **this journal is aimed at all economists, including those not conversant with recent work in the subspecialty of the author.** The goal is to have articles that can be read by 90 percent or more of the AEA membership, as opposed to articles that can only be mastered with abundant time and energy. Articles should be as complex as they need to be, but not more so. Moreover, the necessary complexity should be explained in terms appropriate to an audience presumed to have an understanding of economics generally, but not a specialized knowledge of the author's methods or previous work in this area.

The *Journal of Economic Perspectives* is intended to be scholarly without relying too heavily on mathematical notation or mathematical insights. In some cases, it will be appropriate for an author to offer a mathematical derivation of an economic relationship, but in most cases it will be more important that an author explain why a key formula makes sense and tie it to economic intuition, while leaving the actual derivation to another publication or to an appendix.

JEP does not publish book reviews or literature reviews. Highly mathematical papers, papers exploring issues specific to one non-U.S. country (like the state of agriculture in Ukraine), and papers that address an economic subspecialty in a manner inaccessible to the general AEA membership are not appropriate for the *Journal of Economic Perspectives*. Our stock in trade is original, opinionated perspectives on economic topics that are grounded in frontier scholarship. If you are not familiar with this journal, it is freely available on-line at <http://e-*JEP*.org>.

Guidelines for Preparing *JEP* Proposals

Almost all *JEP* articles begin life as a two- or three-page proposal crafted by the authors. If there is already an existing paper, that paper can be sent to us as a proposal for *JEP*. However, given

the low chances that an unsolicited manuscript will be published in *JEP*, no one should write an unsolicited manuscript intended for the pages of *JEP*. **Indeed, we prefer to receive article proposals rather than completed manuscripts.** The following features of a proposal seek to make the initial review process as productive as possible while minimizing the time burden on prospective authors:

- Outlines should begin with a paragraph or two that precisely states the main thesis of the paper.
- After that overview, an explicit outline structure (I., II., III.) is appreciated.
- The outline should lay out the expository or factual components of the paper and indicate what evidence, models, historical examples, and so on will be used to support the main points of the paper. The more specific this information, the better.
- The outline should provide a conclusion
- Figures or tables that support the article's main points are often extremely helpful.
- The specifics of fonts, formatting, margins, and so forth do not matter at the proposal stage. (This applies for outlines and unsolicited manuscripts).
- Sample proposals for (subsequently) published *JEP* articles are available on request.
- For proposals and manuscripts whose main purpose is to present an original empirical result, please see the specific guidelines for such papers below.

The proposal provides the editors and authors an opportunity to preview the substance and flow of the article. For proposals that appear promising, the editors provide feedback on the substance, focus, and style of the proposed article. After the editors and author(s) have reached agreement on the shape of the article (which may take one or more iterations), the author(s) are given several months to submit a completed first draft by an agreed date. This draft will receive detailed comments from the editors as well as a full set of suggested edits from *JEP*'s Managing Editor. Articles may undergo more than one round of comment and revision prior to publication.

Readers are also welcome to send e-mails suggesting topics for *JEP* articles and symposia and to propose authors for these topics. If the proposed topic is a good fit for *JEP*, the *JEP* editors will work to solicit paper(s) and author(s).

Correspondence regarding possible future articles for *JEP* may be sent (electronically please) to the assistant editor, Ann Norman, at <anorman@JEPjournal.org>. Papers and paper proposals should be sent as Word or pdf e-mail attachments.

Guidelines for Empirical Papers Submitted to *JEP*

The *JEP* is not primarily an outlet for original, frontier empirical contributions; that's what refereed journals are for! Nevertheless, *JEP* occasionally publishes original empirical analyses that appear uniquely suited to the journal. In considering such proposals, the editors apply the following guidelines (in addition to considering the paper's overall suitability):

1) The paper's main topic and question must not already have found fertile soil in refereed journals. *JEP* can serve as a catalyst or incubator for the refereed literature, but it is not a competitor.

2) In addition to being intriguing, the empirical findings must suggest their own explanations. If the hallmark of a weak field journal paper is the juxtaposition of strong claims with weak evidence, a *JEP* paper presenting new empirical findings will combine strong evidence with weak claims. The empirical findings must be robust and thought provoking, but their interpretation should not be portrayed as the definitive word on their subject.

3) The empirical work must meet high standards of transparency. *JEP* strives to only feature new empirical results that are apparent from a scatter plot or a simple table of means. Although *JEP* papers can occasionally include regressions, the main empirical inferences should not be regression-dependent. Findings that are not almost immediately self-evident in tabular or graphic form probably belong in a conventional refereed journal rather than in *JEP*.

Call For Papers and Sessions

2017 American Economic Association Annual Meeting

Chicago, IL January 6–8, 2017

The AEA, in conjunction with 55 associations in related disciplines, assembles over 12,000 of the best minds in economics to network and celebrate new achievements in economic research. This is the premiere event to share your work with colleagues.

Authors are invited to submit all proposals electronically via the American Economic Association website at **http://www.aeaweb.org/Annual_Meeting**

AEA papers covering a wide array of economics topics will be included on the 2017 program. Papers on econometric or mathematical methods should be submitted to the Econometric Society.

- Some of the papers presented at the annual meeting are published in the May ***American Economic Review (the Papers & Proceedings)***.
- All ***submitted*** papers, whether individual or part of a session, must have at least one author who is a member of the AEA. The Association discourages multiple proposals from the same person.
- Please submit complete information. No changes are accepted until a decision is made about inclusion in the program (usually in July). Do not send a complete paper until selections are made.
- Proposals for complete sessions have historically had a higher probability of inclusion than papers submitted individually. Use EconHarmony to form integrated complete sessions consisting of 3–4 papers before the session organizer submits the proposal.

February 1, 2016	EconHarmony Opens on the AEA website
March 1, 2016	AEA Begins Accepting All Proposals
April 1, 2016	Deadline for Individual Paper Proposals
April 15, 2016	Deadline for Complete Session Proposals

American Economic Association
2014 Broadway, Suite 305
Nashville, TN 37203

Phone: (615) 322-2595
Fax: (615) 343-7590
Email: aeainfo@vanderbilt.edu

The American Economic Association

Correspondence relating to advertising, business matters, permission to quote, or change of address should be sent to the AEA business office: aeainfo@vanderbilt.edu. Street address: American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. For membership, subscriptions, or complimentary *JEP* for your e-reader, go to the AEA website: http://www.aeaweb.org. Annual dues for regular membership are $20.00, $30.00, or $40.00, depending on income; for an additional fee, you can receive this journal, or any of the Association's journals, in print. Change of address notice must be received at least six weeks prior to the publication month.

The Journal of

Economic Perspectives

Winter 2016, Volume 30, Number 1

Symposia

The Bretton Woods Institutions

Oil and Gas Markets

Articles

Features

AMERICAN ECONOMIC ASSOCIATION